AD_____

Award Number: DAMD17-96-1-6012

TITLE: Computer-aided Classification of Malignant and Benign Lesions on Mammograms

PRINCIPAL INVESTIGATOR: Berkman Sahiner, M.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, Michigan 48103-1274

REPORT DATE: May 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 2001	Final (1 May 96 - 30 Apr 01)

4. TITLE AND SUBTITLE
Computer-aided Classification of Malignant and Benign Lesions on Mammograms

5. FUNDING NUMBERS
DAMD17-96-1-6012

6. AUTHOR(S)
Berkman Sahiner, M.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Ann Arbor, Michigan 48103-1274

E-Mail: berki@umich.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; Distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

New methods have been investigated for computerized characterization of mammographic masses and microcalcifications. A mass segmentation method based on an active contour model was developed. The resulting segmentation algorithm was shown to be within the inter-observation variation of radiologists' hand segmentation. Morphological, texture, and spiculation features were extracted from the segmented mass and its margins. New classifiers based on statistical methods, genetic algorithms and neural networks were developed. The high-sensitivity classifier developed in this project was shown to have a significantly higher sensitivity than competing classifiers at the same specificity levels. The effect of the mass classification algorithm on radiologists' classification was evaluated using an observer study. It was shown that the radiologists' classification was significantly improved when they were aided by the computerized classification scores. A microcalcification detection algorithm was applied for automated detection of individual microcalcifications with a region of interest. The individual microcalcifications were segmented from the background using an automated algorithm. Morphological and texture features were extracted from computer-detected microcalcifications, and were used in a statistical classifier to distinguish between malignant and benign microcalcifications. Using an observer performance study, it was shown that the developed automated microcalcification characterization method was significantly more accurate than experienced radiologists.

14. SUBJECT TERMS
Breast Cancer, Computer-aided diagnosis

15. NUMBER OF PAGES
243

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

Treatment of the breast cancer at an early stage is the most significant means of improving the survival rate of the patients. Mammography is currently the most sensitive method for detecting early breast cancer, and it is also the most practical for screening. However, the positive predictive value of mammographic diagnosis is only about 15%-30%. As the number of patients who undergo mammography increases, it will be increasingly important to improve the positive predictive value of mammography in order to reduce costs and patient discomfort. In this proposal, our goal was to investigate the problem of classifying mammographic lesions as malignant or benign using computer vision, automatic feature extraction, statistical classification, and artificial intelligence techniques. Our efforts were concentrated on the computer-aided classification of two kinds of breast abnormalities, masses and microcalcifications, which are the primary mammographic signs of malignancy. We investigated computerized extraction of useful features for the differentiation of malignant and benign cases for both abnormalities, and the application of classical statistical classifiers and newly developed paradigms such as neural networks and genetic algorithms for the classification task. Our purposes were to i) improve existing techniques, devise new methods, and identify the preferred approaches for the classification of mammographic lesions, ii) show that computerized classification of mammographic lesions is feasible, and iii) develop a computerized program that can subsequently be shown to improve radiologists' classification of mammographic abnormalities.

Body

Technical objective 1: Database collection

We have digitized over 600 new films from over 150 patients where each case contained either a

biopsy proven mass or a biopsy proven microcalcification cluster. The expert mammographer in this project has been reading these films, and marking the locations of the lesions on the films. He has also rated the visibility and likelihood of malignancy for most of the lesions.

Technical objective 2: Feature extraction for masses

Extraction of the mass shape

We have developed a segmentation method based on k-means clustering [1, 2] followed by an active contour model [3-5], and a spiculation segmentation algorithm [6]. We have also investigated segmentation methods based on Gauss-Markov random fields and based on neural networks [7, 8].

The k-means algorithm classifies a pixel p_i as either an object pixel or a background pixel by clustering the feature vectors F_i for all the pixels in a region of interest (ROI). The algorithm starts by choosing initial cluster center vectors, for the object and the background. Pixels are classified as background or object pixels based on the Euclidean distance between the cluster vector and the cluster center vector. Using this initial classification, two new cluster center vectors are computed. If the cluster centers are different from the previous ones, the procedure of temporary classification is repeated, otherwise, clustering is completed. Other implementation details of and examples of segmentation have been provided in the literature [2].

Although initial mass segmentation resulted in reasonable mass shapes for most of the masses, further refinement was necessary before detection and segmentation of the spiculations. We used an active contour model for mass shape refinement. An active contour is a deformable continuous curve, whose shape is controlled by internal forces (the model, or a-priori knowledge about the object to be segmented) and external forces (the image). The internal energy components in our active contour model are the continuity and curvature of the contour, and the

external energy components are the negative of the smoothed image gradient. The result of the clustering-based segmentation was used to initialize the deformable model. Our method has been described in detail in the literature [6].

This automated method for the segmentation of the central tumor was quantitatively compared with manual segmentation by two expert radiologists (R1 and R2) using three similarity or distance measures on a data set of 100 masses [5]. The inter-observer difference in these measures between the two radiologists was compared to the corresponding differences between the computer and the radiologists. Using three similarity measures and data from two radiologists, a total of six statistical tests were performed. The difference between the computer and the radiologist segmentation was significantly larger than the inter-observer variability in only one test. These results suggest that the segmentation method for outlining the central tumor in our mass classification algorithm is satisfactory [5].

The active contour segmentation results are close to the visually perceived object boundaries, but spiculations are not detected. To automatically outline the spiculations, we used a spiculation detection method, which uses the distribution of the angle between θ two vectors for each border pixel b. The first vector is the gradient direction at a border pixel in a band of pixels around the segmented mass, and the second vector is the direction from this image pixel to the border pixel b [4, 6]. On a data set of 249 mammograms (69 spiculated and 180 non-spiculated), we were able to correctly identify 85% of the spiculated masses and 80% of the non-spiculated masses [6]. In the final stage of our algorithm, the spiculations were appended to the already extracted mass shape. Examples of automatically outlined spiculations are provided in the literature [3, 6].

The rubber-band straightening transform

We have designed a novel image transformation method, referred to as the rubber-band straightening transform (RBST) to map the band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the border of a mass is expected to appear approximately as a horizontal edge, and spiculations are expected appear approximately as vertical lines. The radially oriented features in the original image will therefore become rectilinear in the transformed image. The RBST facilitates the computerized extraction of the important image features. Implementation details and examples can be found in the literature [1, 2]. The effect of the RBST on mass characterization accuracy is discussed under technical objective 5.

Extraction of morphological features

We developed algorithms to extract thirteen morphological features from the segmented masses. The first five morphological features were based on the normalized radial length (NRL), defined as the Euclidean distance from the object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object [6]. In our previous studies, we found that NRL mean, standard deviation, entropy, area ratio, and zero crossing count were useful for discriminating between objects containing masses and normal tissue [9]. The next six features were the perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast of the object. The definition of these features can be found in the literature [9]. The twelfth feature, convexity, was defined as the ratio of the area of the segmented object to the area of the smallest convex object that contained the object. If the object was convex, as was the case with many benign masses, then this feature would approach its maximum value of unity. If the object shape was highly non-convex, as was the case with many malignant masses, then the value of this feature would be small. The last feature was the summary Fourier descriptor measure [6], which was based on the Fourier transform of the object boundary sequence. Objects with

irregular contours have more high-frequency components than those with smooth contours. The Fourier descriptor measure therefore contains potentially useful information for discriminating between benign and malignant masses.

Extraction of texture features

The texture of the region surrounding the mass can yield important features for its classification. Since spiculations and the gradient of the opacity caused by the mass are approximately radially oriented, the texture of the region surrounding a mass is expected to have a radial dependence. However, most texture extraction methods are designed for texture orientations in a uniform direction (horizontal, vertical, or at a certain angle between these two directions). As explained previously, we have designed the RBST to be able to extract meaningful texture features from the region surrounding a mass.

The texture features extracted from the RBST images include 13 texture measures, each calculated at 4 directions and 10 distances, from the spatial gray-level dependence (SGLD) matrices and 20 run-length statistics (RLS) features, as described in our previous work [2]. The definition of these features [10, 11] and the parameters used in this study can be found in the literature [2].

Other features

Three spiculation features are extracted based on whether points on the mass contour lie on the path of a spiculation. Since a spiculation is a linear structure, the image gradients at different points that lie on the same spiculation have similar phase directions. We have defined a spiculation measure in terms of the statistics of these phase directions [6]. Three spiculation features were defined in terms of this spiculation measure [5].

Technical objective 3: Feature extraction for microcalcifications

Extraction of microcalcifications

In an automated microcalcification classification algorithm, the individual microcalcifications within a cluster need to be automatically identified. The area to search for these microcalcifications can be either manually identified, or provided by an automated microcalcification detection algorithm. We therefore developed a technique to detect individual microcalcifications in a given ROI. The size of the ROI was 5.1cm x 5.1 cm. The automated detection algorithm is based on our previous work [12]. First, a difference-image technique is applied to the ROI for signal-to-noise ratio enhancement [13]. Then, the gray level histogram of the enhance ROI is determined and global and local gray level thresholding are used to locate potential signal locations [14].

Extraction of morphological features

Starting from these locations, and automated region growing technique extracted the signal location as the connected pixels above a gray-level threshold, which was determined as the product of the local root-mean-square noise and an input SNR threshold. After initial experimentation, an SNR threshold of 2.0 was chosen for all cases. Five features, namely the area, mean density, eccentricity, moment ratio, and area ratio were defined in terms of the first and second moments of the extracted microcalcification signals. Since the variations of the shapes and sizes of the individual microcalcifications within a cluster are important for microcalcification classification, the maximum, mean, standard deviation, and coefficient of variation of these individual features were computed for each cluster. Another feature describing the number of microcalcifications was also added, resulting in a 21-dimensional morphological feature space [15].

Extraction of texture features

The low-frequency background was subtracted using a background subtraction technique [16]. After background correction, four gray level difference statistics (GLDS) features, namely mean, entropy, contrast, and angular second moment were extracted at four different directions from the ROI containing the microcalcification cluster [17]. SGLD texture matrices were computed as discussed in our original proposal. Forty SGLD matrices were derived from each ROI at different 10 distances and 4 directions. Thirteen features were extracted from each SGLD matrix, and features extracted at axial directions and diagonal directions were averaged. The final texture space therefore contained 260 SGLD features and 21 GLDS features for each ROI.

Technical objective 4: Development of classifiers

Linear discriminant analysis with stepwise feature selection

For classification of lesions as malignant or benign, we have implemented Fisher's linear discriminant with stepwise feature selection. For a two-class problem, Fisher's linear discriminant projects the multi-dimensional feature space onto the real line in such a way that the ratio of between-class sum of squares to within-class squares is maximized after the projection. This is the optimal classifier if the two classes have a multivariate Gaussian distribution with equal covariance matrices. When the data size is limited, the inclusion of inappropriate features may reduce the test accuracy. In such a case, feature selection becomes necessary. We have used stepwise feature selection to reduce the dimensionality of the feature space before Fisher's linear discriminant is applied [2].

The understanding of the performance of the classifier designed with different schemes will allow us to utilize a limited sample set efficiently. The relationship between the bias of the accuracy estimate of a classifier, the number of available samples, the number of features selected

in stepwise feature selection, and the estimation of coefficients for Fisher's linear discriminant were studied as part of this project [18]. Our results indicated that the resubstitution estimate was always optimistically biased, except in cases where the parameters of stepwise feature selection were chosen such that too few features were selected by the stepwise procedure. When feature selection was performed using only the design samples, the hold-out estimate was always pessimistically biased. When feature selection was performed using the entire finite sample space, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, the number of available samples, and their statistical distribution [18].

Neural networks

In classification of mammographic lesions, the cost of missing a malignant case is much larger than that of misclassifying a normal case. The decision threshold therefore cannot be determined without a well-designed cost-benefit analysis. Receiver Operating Characteristic (ROC) analysis is a commonly-used methodology for representing the tradeoff between the true-positive fraction (TPF) and the false-positive fraction (FPF) in a two-group classification task. The area A_{TPF_0} above a sensitivity level TPF_0 under the ROC curve represents the average specificity above that sensitivity level. By maximizing A_{TPF_0}, where TPF_0 is close to 1, one can design a classifier that has good specificity at high sensitivity. In this project, we tried to develop a methodology for training a backpropagation neural network (BPN) by maximizing this criterion [7, 19].

To test our new BPN training algorithm, we used a randomly-generated Gaussian data set. Our results indicated that for small number of hidden-layer nodes, the new training algorithm slightly decreased the false-positives for a TPF of 0.8 and above. When the number of

hidden-layer nodes was increased, the difference between the two training algorithms diminished. This simulation study therefore demonstrated that the new training algorithm would be useful only if the number of hidden layer nodes is small [19].

Development of a hierarchical classifier

A hierarchical classifier, which combines an unsupervised adaptive resonance network (ART2) and a supervised linear discriminant classifier (LDA) was developed for the classification of mammographic masses as malignant or benign [20]. At the first stage, the ART2 network separated the masses based on the similarity of the input vectors. At the second stage, a separate LDA model was formulated within each class to classify the masses as malignant or benign. The ART2 network was presented with texture features extracted from RBST images, as described previously. The ART2 network classified the masses two classes: one containing only malignant masses and the other containing a mix of malignant and benign masses. The masses from the malignant class were classified by ART2. The masses from the mixed class were input to a supervised linear discriminant classifier (LDA). In this way, some malignant masses were separated and classified by ART2 and less distinguishable masses were classified by the LDA. For the evaluation of classifier performance, 348 regions of interest (ROIs) were used. The area under the ROC curve was 0.81 for the hybrid classifier, 0.89 for a backpropagation neural network (BPN), and 0.78 for the LDA. These result indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.

BPNs for microcalcification and mass classification

The texture and morphological features described in Technical Objectives 2 and 3 were used in a backpropagation neural network (BPN) for classification of masses [4] and

microcalcification clusters [21], respectively. First, stepwise feature selection was used to select effective features for classification. Then, several BPN structures were tested their classification accuracy. The BPNs employed a modified delta-bar-delta rule to improve the convergence rate [16].

For the classification of microcalcifications number of hidden nodes in the BPNs varied between 1 and 10. The classifier was tested on a database of 86 mammograms from 54 cases [21]. The area Az under the ROC curve obtained with different BPN structures varied between 0.88 and 0.86 with the best feature set. An analysis of the dependence of the classification accuracy on BPN architecture indicated that the BPN with one hidden node provided the best classification accuracy. Since a BPN with a single hidden node is equivalent to a linear classifier, this result appears to indicate that a linear classifier may be optimal with this data set and training samples. However, it has to be emphasized that this observation may not apply when the classifiers are trained with large number of samples. The reduction in classification accuracy with increased number of hidden layer nodes in our current study could have been caused by overtraining with a small sample size.

For the classification of masses, a BPN with four input nodes, two hidden-layer nodes, and a single output node was trained using the training set, which consisted of 243 mammograms (116 benign and 127 malignant) from 101 patients. The accuracy of the designed classifier was evaluated by applying the classifier to test cases that had not been used for training. The test data set consisted of 95 mammograms (42 benign and 53 malignant) from 45 patients. A single view was available for nine of these 45 patients. For the remaining 36 test patients, two or more views were available. The case-based classification A_z values for 0.95 for the training set and 0.87 for the test set.

Genetic algorithms

A genetic algorithm (GA)-based feature selection method was developed for the design of high-sensitivity classifiers, which were tailored to yield high sensitivity with high specificity. The fitness function of the GA was based on the ROC partial area index, which is defined as the average specificity above a given sensitivity threshold. The designed GA evolved towards the selection of feature combinations which yielded high specificity in the high sensitivity region of the ROC curve, regardless of the performance at low sensitivity. This is a desirable quality of a classifier used for breast lesion characterization, since the focus in breast lesion characterization is to correctly diagnose as many benign lesions as possible without missing malignancies. The high-sensitivity classifier, formulated as the Fisher's linear discriminant using GA-selected feature variables, was employed to classify 255 biopsy-proven mammographic masses as malignant or benign. SGLD and RLS textures features were extracted from the RBST images, as described previously. The classification accuracy of the high-sensitivity classifier was compared to that of linear discriminant analysis with stepwise feature selection (LDA_{sfs}). With proper GA training, the ROC partial area of the high-sensitivity classifier above a true-positive fraction of 0.95 was significantly larger than that of LDA_{sfs}, although the latter provided a higher total area (A_z) under the ROC curve. By setting an appropriate decision threshold, the high-sensitivity classifier and LDA_{sfs} correctly identified 61% and 34% of the benign masses, respectively, without missing any malignant masses. Our results show that the choice of the feature selection technique is important in computer aided diagnosis, and that the GA may be a useful tool to design classifiers for lesion characterization [22].

Technical objective 5: Evaluation of classification methods

Effectiveness of the computer classifiers

For the task of classifying between malignant and benign masses, we have found that the most effective classifier is an LDA classifier that uses texture features extracted from RBST images, and morphological features extracted from automatically extracted mass shapes [6]. The classification accuracy with features extracted from the RBST images was significantly higher than that using the original ROIs [2]. A data set containing 249 films from 102 patients was used in a leave-one-case-out data partitioning scheme to train and test the classifier. When the leave-one-case-out method was applied to partition the data set into trainers and testers, the average test A_z for the task of classifying the mass on a single mammographic view was 0.83±0.02, 0.84±0.02, and 0.87±0.02 in the morphological, texture, and combined feature spaces, respectively. For classifying a mass as malignant or benign, we combined the leave-one-case-out discriminant scores from different views of a mass to obtain a summary score. In this task, the test A_z value using the combined feature space was 0.91±0.02. Our data set contained 26 prior mammograms and 223 preoperative masses. When only preoperative masses were considered, the case-based A_z values was 0.94.

For the task of classifying between malignant and benign microcalcifications, we have found that the most effective classifier is an LDA classifier that uses texture and morphological features, selected by the stepwise feature selection method [23]. The data set for computerized classification consisted of 112 pairs (CC and MLO or CC and LAT) of mammograms. The number of malignant and benign pairs were 40 and 72, respectively. The scores from the two views of a pair were averaged to obtain a score for the pair. Computer classification scores were analyzed by ROC analysis. The accuracy of the classifier was evaluated by the area A_z and the partial area index $A_z(TPF_0)$ above a true-positive fraction of TPF_0=0.90. The computer classifier had an ROC area of 0.83 and a partial area index of 0.42.

Comparison to radiologist's accuracy

The effect of our mass classification algorithm on radiologist' classification was evaluated using an observer study. Of the 255 films that were used in our previous studies, 15 were used for training the radiologists to use the computer estimation for malignancy, and the remaining 240 were used for the actual evaluation. Six MQSA-approved radiologists assessed the probability of malignancy of the masses with and without CAD. Two experiments, one with single view and another with two views were conducted.

The computer classifier alone distinguished the malignant and benign masses with a test Az of 0.92. The radiologists' Az ranged from 0.78 to 0.91 without CAD and were improved to 0.91 to 0.97 with CAD. For a subset of 77 matched paired views, the radiologists' Az ranged from 0.88 to 0.95 without CAD and were improved to 0.93 to 0.97 with CAD. The improvements were statistically significant with p=0.02 and 0.01, respectively. The average observer ROC curves with and without CAD are shown in Figure 1.

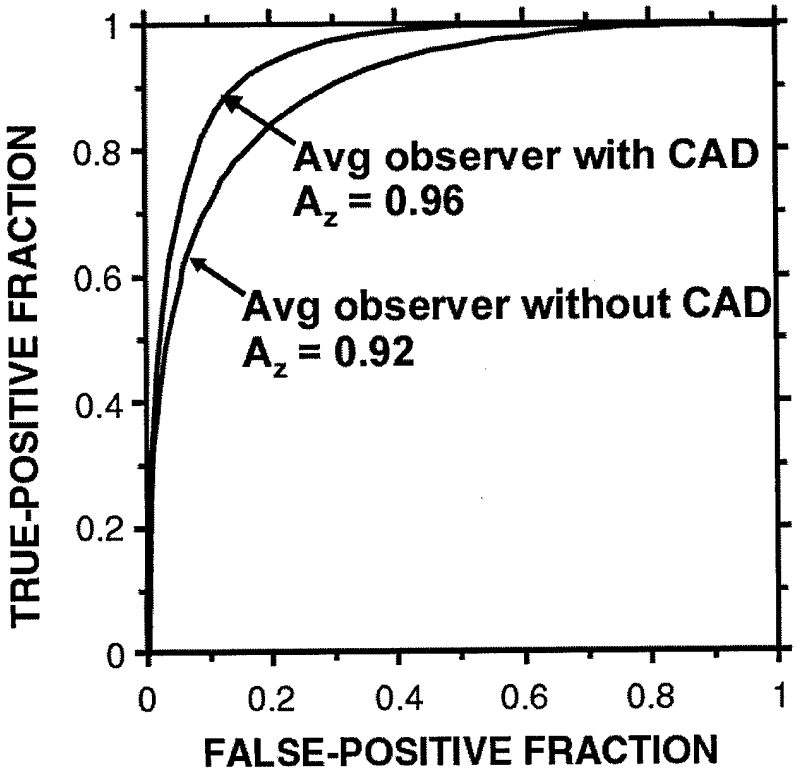


Figure 1. The average ROC curve of radiologists for the two-view classification of mammographic masses as malignant or benign with and without CAD.

The effectiveness of our microcalcification classification algorithm was also evaluated using an observer study in which 7 MQSA-approved radiologists read the same 112 pairs of ROIs described previously. The ROIs were printed on film with a laser printer. The radiologists rated the likelihood of malignancy of each pair on a 10-point rating scale. The case order was randomized for each radiologist. The average ROC curve of 7 radiologists was computed by averaging the slope and intercept parameters of individual ROC curves. The classification accuracy of the radiologists was compared to that of the computer. It was found that the A_z value of the computer was higher than that of all radiologists, and the difference was statistically significant for three of the radiologists ($p=0.03$). The average ROC curve of the 7 radiologists and

the ROC curve of the computer classifier are shown in Figure 2.

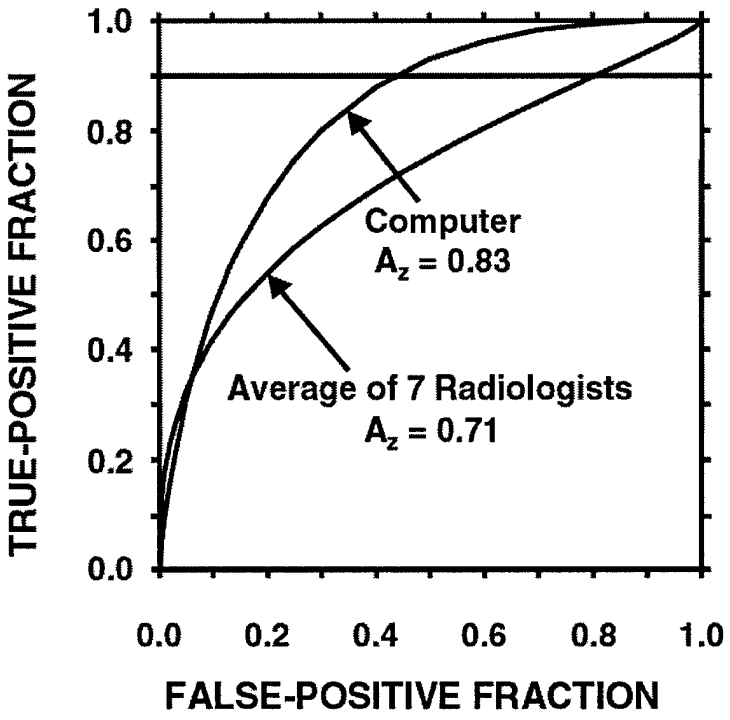


Figure 2. The average ROC curve of the 7 radiologists for the classification of mammographic microcalcifications a malignant or benign and the ROC curve of the computer classifier.

Key Research Accomplishments

- Three new methods (clustering, active contours, and NN-based segmentation) have been investigated for the segmentation of masses on mammograms. It has been shown that the combination of clustering and active contours results in a segmentation that is within the inter-observation variation of radiologists' hand segmentation.
- The rubber-band straightening transform (RBST) has been developed for assisting in feature extraction from mass margins. It has been shown that features extracted from the RBST images are significantly more effective than those extracted from original images.

- A number of morphological, texture and spiculation features have been developed for classification of mammographic masses.
- A high-sensitivity classifier, based on a genetic algorithm for feature selection, has been developed for obtaining better specificity at a high sensitivity for malignancies. It has been shown that the high-sensitivity classifier achieves a significant improvement over stepwise feature selection.
- A hierarchical classifier which combines an unsupervised adaptive resonance network (ART2) and a supervised linear discriminant classifier (LDA) was developed for the classification of mammographic masses as malignant or benign.
- The effect of the mass classification algorithm on radiologists' classification was evaluated using an observer study. Using a database of 240 mammograms, it was shown that the radiologists' classification was significantly improved when they were aided by the computerized classification scores.
- The generalizability of our mass classification method was tested by applying a trained classifier to an independent data set containing biopsied masses.
- A previously-existing microcalcification detection algorithm was applied for automated detection of individual microcalcifications with a region of interest
- The individual microcalcifications were segmented from the background using an automated algorithm
- Morphological and texture features were extracted from computer-detected microcalcifications.
- Texture features extracted from a region of interest containing the microcalcifications were used in a backpropagation neural network for classification of microcalcifications as

malignant and benign.

- The classification accuracy of the morphological features was evaluated by using the morphological feature space alone and by combining the morphological and texture feature spaces.
- Using an observer performance study, it was shown that the developed automated microcalcification characterization method was significantly more accurate than experienced radiologists.

Reportable Outcomes

Journal Papers

1. H.P. Chan, B. Sahiner, N. Petrick, M.A. Helvie, K.L. Lam, D.D. Adler, and M.M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," *Physics of Medicine and Biology,* 1997, 42:549-567.
2. B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, and M.M. Goodsitt, "Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis," *Medical Physics,* 1998, 25:516-526.
3. H.P. Chan, B. Sahiner, K.L. Lam, N. Petrick, M.A. Helvie, M.M. Goodsitt, and D.D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," *Medical Physics,* 1998, 25:2007-2019.
4. B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, and M.M. Goodsitt, "Design of a high-sensitivity classifier based on a genetic algorithm: Application to computer-aided diagnosis," *Physics of Medicine and Biology,* 1998, 43:2853-2871.

5. H.P. Chan, B. Sahiner, M.A. Helvie, N. Petrick, M.A. Roubidoux, T.E. Wilson, D.D. Adler, C. Paramagul, J.S. Newman, S.S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," *Radiology*, 1999, 212:817-827.

6. L.M. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," *IEEE Trans. Medical Imaging*, 1999, 18: 1178-1187.

7. B. Sahiner, H.P. Chan, N. Petrick, R.F. Wagner, L.M. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size," *Medical Physics,* 2000, 27: 1509-1522.

8. H-P. Chan, M.A. Helvie, N. Petrick, B. Sahiner, D.D. Adler, C. Paramagul, M.A. Roubidoux, C.E. Blane, L.K. Joynt, T.E. Wilson, L.M. Hadjiiski, M.M. Goodsitt, "Digital mammography: Observer performance study of effects of pixel size on radiologists' characterization of malignant and benign microcalcifications," *Academic Radiology*, 2001, 8:454-466.

9. B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, L.M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics* (in press), 2001.

10. B. Sahiner, N. Petrick, H.P. Chan, L.M. Hadjiiski, C. Paramagul, M.A. Helvie, M.N. Gurcan, "Computer-Aided Characterization of Mammographic Masses: Accuracy of Mass Segmentation and its Effects on Characterization,", *IEEE Trans. Medical Imaging (submitted)*, 2001.

Conference Proceedings

1. B. Sahiner, H.P. Chan, N. Petrick, M.M. Goodsitt, and M.A. Helvie, "Characterization of masses on mammograms: Significance of the use of the rubber band straightening transform," *Proc. SPIE Medical Imaging 1997*, 3034:491-500.

2. B. Sahiner, H.P. Chan, N. Petrick, S.S. Gopal, and M.M. Goodsitt, "Neural network design for optimization of the partial area under the receiver operating characteristic curve," *Proc. IEEE International Conference on Neural Networks 1997,* 4:2468-2471.

3. S.S. Gopal, B. Sahiner, H.P. Chan, and N. Petrick, "Neural network based segmentation using *a priori* image models," *Proc. IEEE International Conference on Neural Networks 1997,* 4:2455-2459.

4. B. Sahiner, H.P. Chan, N. Petrick, R.F. Wagner, and L.M. Hadjiiski, "Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size," *Proc. SPIE Medical Imaging, 1999,* 3661:499-510.

5. L.M. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, "Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms," *Proc. SPIE Medical Imaging, 1999,* 3661:464-473.

6. B. Sahiner, H-P. Chan, N. Petrick, L.M. Hadjiiski, M.A. Helvie, S. Paquerault, "Active contour models for segmentation and characterization of mammographic masses," to appear in proceedings of International Workshop on Digital Mammography, Toronto, June 2000.

7. L.M. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, M.N. Gurcan, "Analysis of temporal change of mammographic features for computer-aided characterization of malignant and benign masses," *in Proceedings of SPIE Medical Imaging (in Press),* 4322, 2001.

Meeting Abstracts

1. S. Sanjay-Gopal, H.P. Chan, B. Sahiner, N. Petrick and M.A. Helvie, "Evaluation of automated methods for the segmentation of lesion boundaries in mammograms for computer aided diagnosis," *83rd Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, December 1997.
2. H.P. Chan, B. Sahiner, M.A. Helvie, C. Paramagul, J. Newman, S. Sanjay-Gopal, N. Petrick, D.D. Adler, M. Roubidoux and T. Wilson, "Effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses on mammograms: an ROC study," *83rd Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, December 1997.
3. B. Sahiner, H.P. Chan, M.A. Helvie, T.E. Wilson, S. Sanjay-Gopal, N. Petrick, "Computerized classification of mammographic masses using morphological features," *84th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, Nov. 1998.
4. L.M. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, M.M. Goodsitt, "Characterization of malignant and benign masses on mammograms based on a hierarchical classifier," *84th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, Nov. 1998.

5. H.P. Chan, B. Sahiner, M.A. Helvie, N. Petrick, L.M. Hadjiiski, M.A. Roubidoux, "Computer-aided breast cancer diagnosis: Comparison of computerized classification with radiologists' performance," *85th Scientific Assembly and Annual Meeting of the Radiological Society of North America*, Chicago, IL, Nov. 1999.

6. B. Sahiner, H.P. Chan, N. Petrick, L.M. Hadjiiski, M.A. Helvie, "Computer-aided characterization of mammographic lesions," *Era of Hope Meeting*, U. S. Army Medical Research and Materiel Command, Department of Defense, Breast Cancer Research Program, Atlanta, Georgia, June 8-12, 2000.

7. L.M. Hadjiiski, H.P. Chan, B. Sahiner, N. Petrick, M.A. Helvie, M.N. Gurcan, "Computer-aided classification of malignant and benign breast masses by analysis of interval change of features in temporal pairs of mammograms," *Radiology*, 217(P):435, 2000.

List of personnel receiving pay from the research effort: Berkman Sahiner, Ph.D.

Conclusions

New methods have been investigated for computerized characterization of mammographic masses and microcalcifications as malignant or benign. These methods included algorithms for segmentation of masses and microcalcifications from the background in a region of interest, texture and morphological feature extraction techniques, and classifiers based on statistical methods, genetic algorithms and neural networks. The classification accuracy of the automated algorithms were compared to the accuracy of radiologists experienced in mammographic interpretation. It has been shown that the accuracy of the microcalcification classification

algorithm is significantly better than that of radiologists. It has also been shown that the mass classification algorithm can significantly improve the characterization accuracy of radiologists when they interpret the mammograms with the aid of the developed algorithms. These results are very encouraging for the clinical implementation of computer-aided lesion characterization in mammography. Besides clinical implementation, future work includes the use of prior mammograms and breast ultrasound for further improvement of computerized lesion characterization.

References

[1] B. Sahiner, "USAMRMC Annual Report, Year 1," 1997.

[2] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Med. Phys.*, vol. 25, pp. 516-526, 1998.

[3] B. Sahiner, "USAMRMC Annual Report, Year 3," 1999.

[4] B. Sahiner, H. P. Chan, N. Petrick, L. M. Hadjiiski, M. A. Helvie, and S. Paquerault, "Active contour models for segmentation and characterization of mammographic masses," presented at The 5th International Workshop on Digital Mammography, Toronto, Canada, 2000.

[5] B. Sahiner, N. Petrick, H. P. Chan, L. M. Hadjiiski, C. Paramagul, M. A. Helvie, and M. N. Gurcan, "Computer-Aided Characterization of Mammographic Masses: Accuracy of Mass Segmentation and its Effects on Characterization," *IEEE Trans. Med. Img.*, 2001 (submitted).

[6] B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of

mammographic mass characterization using spiculation measures and morphological features," *Med. Phys.*, 2001 (in press).

[7] B. Sahiner, "USAMRMC Annual Report, Year 2," 1998.

[8] S. S. Gopal, B. Sahiner, H.-P. Chan, and N. Petrick, "Neural network based segmentation using a priori image models," *Proc. IEEE Int. Conf. Neural Net.*, vol. 4, pp. 2455-2459, 1997.

[9] N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," *Med. Phys.*, vol. 26, pp. 1642-1654, 1999.

[10] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Sys. Man. and Cybern.*, vol. SMC-3, pp. 610-621, 1973.

[11] M. M. Galloway, "Texture classification using gray level run lengths," *Comp. Graph. Img Proc.*, vol. 4, pp. 172-179, 1975.

[12] H. P. Chan, S. C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," *Med Phys*, vol. 22, pp. 1555-1567, 1995.

[13] H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," *Invest. Radiol.*, vol. 25, pp. 1102-1110, 1990.

[14] H. P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon, and P. M. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in mammography," *Med Phys*, vol. 14,

pp. 538-548, 1987.

[15] H. P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature space," *Med. Phys.*, vol. 25, pp. 2007-2019, 1998.

[16] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," *IEEE Trans. Med. Img.*, vol. 15, pp. 598-610, 1996.

[17] H. P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, L. M. Hadjiiski, and M. A. Roubidoux, "Computer-aided breast cancer diagnosis: Comparison of computerized classification with radiologists' performance," *Radiol.*, vol. 213(P), pp. 322-323, 1999.

[18] B. Sahiner, H. P. Chan, N. Petrick, R. F. Wagner, and L. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size.," *Med. Phys.*, vol. 27, pp. 1509-1522, 2000.

[19] B. Sahiner, H. P. Chan, N. Petrick, S. S. Gopal, and M. M. Goodsitt, "Neural network design for optimization of the partial area under the receiver operating characteristic curve," presented at IEEE International Conference on Neural Network, Houston, TX, 1997.

[20] L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, and M. A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," *IEEE Trans. Med. Img.*, vol. 18, pp. 1178-1187, 1999.

[21] H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Leung, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," *Phys. Med. Biol.*, vol.

42, pp. 549-567, 1997.

[22] B. Sahiner, H. Chan, N. Petrick, M. Helvie, and M. Goodsitt, "Design of a high-sensitivity classifier based on a genetic algorithm: Application to computer-aided diagnosis," *Phys. Med. Biol.*, vol. 43, pp. 2853-2871, 1998.

[23] B. Sahiner, "USAMRMC Annual Report, Year 4," 2000.

Appendices

1. H.P. Chan, B. Sahiner, N. Petrick, M.A. Helvie, K.L. Lam, D.D. Adler, and M.M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," *Physics of Medicine and Biology,* 1997, 42:549-567.
2. B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, and M.M. Goodsitt, "Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis," *Medical Physics,* 1998, 25:516-526.
3. H.P. Chan, B. Sahiner, K.L. Lam, N. Petrick, M.A. Helvie, M.M. Goodsitt, and D.D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," *Medical Physics,* 1998, 25:2007-2019.
4. B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, and M.M. Goodsitt, "Design of a high-sensitivity classifier based on a genetic algorithm: Application to computer-aided diagnosis," *Physics of Medicine and Biology,* 1998, 43:2853-2871.

5. H.P. Chan, B. Sahiner, M.A. Helvie, N. Petrick, M.A. Roubidoux, T.E. Wilson, D.D. Adler, C. Paramagul, J.S. Newman, S.S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," *Radiology*, 1999, 212:817-827.

6. L.M. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," *IEEE Trans. Medical Imaging*, 1999, 18: 1178-1187.

7. B. Sahiner, H.P. Chan, N. Petrick, R.F. Wagner, L.M. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size," *Medical Physics,* 2000, 27: 1509-1522.

8. H-P. Chan, M.A. Helvie, N. Petrick, B. Sahiner, D.D. Adler, C. Paramagul, M.A. Roubidoux, C.E. Blane, L.K. Joynt, T.E. Wilson, L.M. Hadjiiski, M.M. Goodsitt, "Digital mammography: Observer performance study of effects of pixel size on radiologists' characterization of malignant and benign microcalcifications," *Academic Radiology*, 2001, 8:454-466.

9. B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, L.M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics* (in press), 2001.

10. B. Sahiner, N. Petrick, H.P. Chan, L.M. Hadjiiski, C. Paramagul, M.A. Helvie, M.N. Gurcan, "Computer-Aided Characterization of Mammographic Masses: Accuracy of Mass Segmentation and its Effects on Characterization,", *IEEE Trans. Medical Imaging (submitted)*, 2001.

11. B. Sahiner, H.P. Chan, N. Petrick, M.M. Goodsitt, and M.A. Helvie, "Characterization of masses on mammograms: Significance of the use of the rubber band straightening transform," *Proc. SPIE Medical Imaging 1997*, 3034:491-500.

12. B. Sahiner, H.P. Chan, N. Petrick, S.S. Gopal, and M.M. Goodsitt, "Neural network design for optimization of the partial area under the receiver operating characteristic curve," *Proc. IEEE International Conference on Neural Networks 1997,* 4:2468-2471.

13. S.S. Gopal, B. Sahiner, H.P. Chan, and N. Petrick, "Neural network based segmentation using *a priori* image models," *Proc. IEEE International Conference on Neural Networks 1997,* 4:2455-2459.

14. B. Sahiner, H.P. Chan, N. Petrick, R.F. Wagner, and L.M. Hadjiiski, "Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size," *Proc. SPIE Medical Imaging, 1999,* 3661:499-510.

15. L.M. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, "Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms," *Proc. SPIE Medical Imaging, 1999,* 3661:464-473.

16. B. Sahiner, H-P. Chan, N. Petrick, L.M. Hadjiiski, M.A. Helvie, S. Paquerault, "Active contour models for segmentation and characterization of mammographic masses," to appear in proceedings of International Workshop on Digital Mammography, Toronto, June 2000.

17. L.M. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, M.N. Gurcan, "Analysis of temporal change of mammographic features for computer-aided characterization of malignant and benign masses," *in Proceedings of SPIE Medical Imaging (in Press),* 4322, 2001.

Phys. Med. Biol. 42 (1997) 549–567. Printed in the UK PII: S0031-9155(97)76676-4

Appendix 1

Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network

Heang-Ping Chan†, Berkman Sahiner, Nicholas Petrick, Mark A Helvie, Kwok Leung Lam, Dorit D Adler and Mitchell M Goodsitt

Department of Radiology, University of Michigan, Ann Arbor, MI, USA

Received 22 July 1996

Abstract. We investigated the feasibility of using texture features extracted from mammograms to predict whether the presence of microcalcifications is associated with malignant or benign pathology. Eighty-six mammograms from 54 cases (26 benign and 28 malignant) were used as case samples. All lesions had been recommended for surgical biopsy by specialists in breast imaging. A region of interest (ROI) containing the microcalcifications was first corrected for the low-frequency background density variation. Spatial grey level dependence (SGLD) matrices at ten different pixel distances in both the axial and diagonal directions were constructed from the background-corrected ROI. Thirteen texture measures were extracted from each SGLD matrix. Using a stepwise feature selection technique, which maximized the separation of the two class distributions, subsets of texture features were selected from the multi-dimensional feature space. A backpropagation artificial neural network (ANN) classifier was trained and tested with a leave-one-case-out method to recognize the malignant or benign microcalcification clusters. The performance of the ANN was analysed with receiver operating characteristic (ROC) methodology. It was found that a subset of six texture features provided the highest classification accuracy among the feature sets studied. The ANN classifier achieved an area under the ROC curve of 0.88. By setting an appropriate decision threshold, 11 of the 28 benign cases were correctly identified (39% specificity) without missing any malignant cases (100% sensitivity) for patients who had undergone biopsy. This preliminary result indicates that computerized texture analysis can extract mammographic information that is not apparent by visual inspection. The computer-extracted texture information may be used to assist in mammographic interpretation, with the potential to reduce biopsies of benign cases and improve the positive predictive value of mammography.

1. Introduction

Mammography is the most sensitive method for detection of early breast cancer. However, the specificity for classification of malignant and benign lesions from mammographic images is quite low. In the United States, the positive predictive value, i.e., the ratio of the number of breast cancers found to the total number of biopsies, of mammography is typically between 15 and 30% (Kopans 1991, Adler and Helvie 1992). An improvement in the positive predictive value would reduce health care costs and eliminate the anxiety and morbidity of patients who would have to undergo unnecessary biopsy otherwise. One

† Address for correspondence: Heang-Ping Chan, PhD, Department of Radiology, University of Michigan Hospital, 1500 E Medical Center Drive, 2910 Taubman Center, Ann Arbor, MI 48109-0326, USA. E-mail address: chanhp@umich.edu

0031-9155/97/030549+19$19.50

of the potential approaches to improving the specificity of mammography is the use of computerized feature extraction techniques to extract information that may not be readily perceived by human readers. The computer-extracted features may complement the visual characteristics of the mammographic abnormalities and provide additional information to the radiologists in distinguishing malignant and benign lesions. The computer-extracted features, alone or in combination with human-perceived features, may also be input to a trained classifier to estimate the likelihood of malignancy of a mammographic lesion, thereby assisting radiologists in making diagnostic decisions.

A number of researchers have attempted to develop feature extraction and classification techniques for masses (Ackerman and Gose 1972; Kilday *et al* 1993, Huo *et al* 1995, Sahiner *et al* 1996a) or microcalcifications (Wee *et al* 1975, Fox *et al* 1980, Chan *et al* 1992, Chitre *et al* 1993, Chan *et al* 1994b, Shen *et al* 1994, Chan *et al* 1995a, c, d, Wu *et al* 1995, Jiang *et al* 1996, Thiele *et al* 1996). Other researchers used radiologists' ratings of mammographic features or encoded the radiologists' readings with numerical values as input to classifiers (Ackerman *et al* 1973, Gale *et al* 1987, Getty *et al* 1988, D'Orsi *et al* 1992, Wu *et al* 1993, Baker *et al* 1996). While the accuracy of lesion characterization in these studies varied, they demonstrated that computer-aided classification has the potential to improve the malignant and benign diagnosis of breast lesions. We have been developing computerized feature-extraction techniques for classification of masses or microcalcifications (Chan *et al* 1992, 1994b, 1995a, c, d, Sahiner *et al* 1996a). The extracted features are analysed by linear or non-linear classifiers which are trained for a specific classification task. We have found that texture features are effective for differentiation of masses and normal tissues (Chan *et al* 1995b, Wei *et al* 1995b), and that morphological features can be used to distinguish malignant and benign clustered microcalcifications (Chan *et al* 1995c). Because the tissue texture in regions containing microcalcifications associated with a malignant process may be different from that associated with a benign process, in the present study we analysed texture features from a region of interest (ROI) containing clustered microcalcifications (Chan *et al* 1995d). The effectiveness of these texture features, in combination with a backpropagation neural network classifier (Freeman and Skapura 1991), for the differentiation of malignant and benign microcalcifications was evaluated. The performance of the neural network was analysed with receiver operating characteristic (ROC) methodology (Swets and Pickett 1982, Metz *et al* 1990).

2. Materials and methods

2.1. Case selection and digitization

In this study, 86 mammograms with clustered microcalcifications were selected from patient files in the Department of Radiology at the University of Michigan. The mammograms were acquired with dedicated mammographic systems with a 0.3 mm focal spot, molybdenum (Mo) anode and 0.03 mm Mo filter. A Kodak Min R/MRE mammographic screen–film system using extended cycle processing was employed as the image receptor. The selection criteria were that the mammogram contained a cluster of microcalcifications, that about half of the case samples were malignant and half were benign, and that no grid lines were visible on the mammogram. The data set included 86 films, some of which were films of different views from the same patient. A total of 54 different patients were included in the data set. There were 41 malignant (26 patients) and 45 benign (28 patients) clusters. The malignant and benign pathology of the microcalcifications had been proven by open surgical biopsy and histologic analysis. The visibility of the microcalcification clusters was ranked

by experienced radiologists on a scale of 1–5 (1, very obvious; 5, very subtle) relative to the range of cases seen in clinical practice. The histogram of the visibility for the 86 clusters is shown in figure 1.

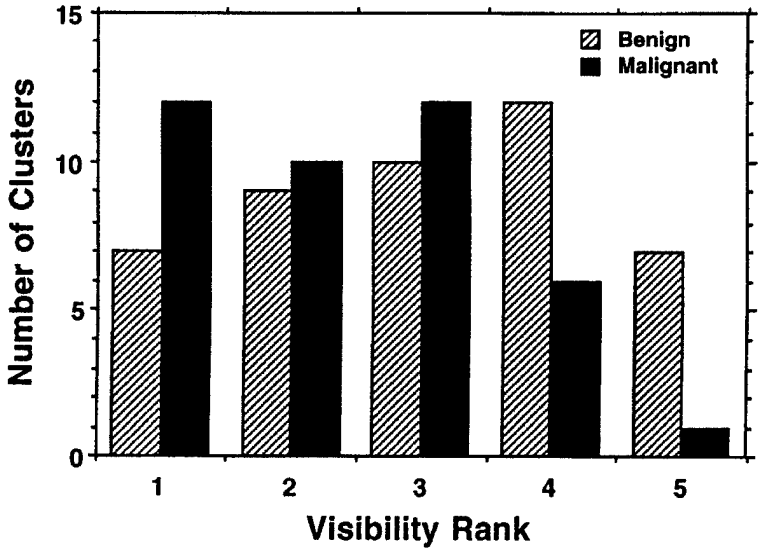


Figure 1. A histogram of the subjective ranking of the visibility of the 86 microcalcification clusters on the mammograms. The clusters were ranked on a five-point scale relative to the range of visibility of clusters found in clinical practice (1, very obvious; 5, very subtle).

All mammograms were digitized with a laser film scanner (Lumisys DIS-1000) at a pixel size of 35 μm × 35 μm and with a 12-bit grey level. The light transmitted through the film was amplified logarithmically before analogue-to-digital conversion. The digitizer had an optical density range of 0–3.5. It was calibrated so that the optical density (O.D.) on film was linearly proportional to the output pixel value in the range of about 0.1–2.8 O.D. with a slope of 0.001 O.D./pixel value. The slope of the calibration curve outside this range decreased gradually. Before input to the detection program, the pixel values were linearly converted such that low optical densities were represented by high pixel values.

In this study, the locations of the microcalcification cluster on each mammogram were identified by radiologists so that only true microcalcification clusters were analysed. An ROI of 1024 × 1024 pixels (corresponding to 3.58 cm × 3.58 cm on the film), with the cluster approximately at its centre was extracted for analysis. This ROI size could enclose the majority of the clusters in the data set. A few of the obvious clusters scattered over a larger area, but the main area of the clusters was covered within the ROI.

The low-frequency background grey levels of each ROI depend mainly on the density of the overlapping breast tissue and the x-ray exposure conditions. The background levels therefore do not relate directly to the presence of the microcalcifications, but they bias the numerical values of the texture features. In order to eliminate the variability in the texture feature distributions caused by these factors that are not related to malignancy, we applied a background correction technique to the ROI before texture feature extraction. This technique has been described in detail previously (Chan *et al* 1995b). Briefly, the grey level at a given pixel of the low-frequency background was estimated as the average of the distance-weighted grey levels of four pixels at the intersections of the normals from the

given pixel to the four edges of the ROI. An example of an original ROI with a malignant cluster, its estimated background image, and the background-corrected ROI is shown in figure 2(a)–(c), respectively. It can be seen that the sloped background grey level of the ROI was removed by the correction. The high-frequency information in the ROI was basically unchanged because the background image only contained low spatial frequencies.

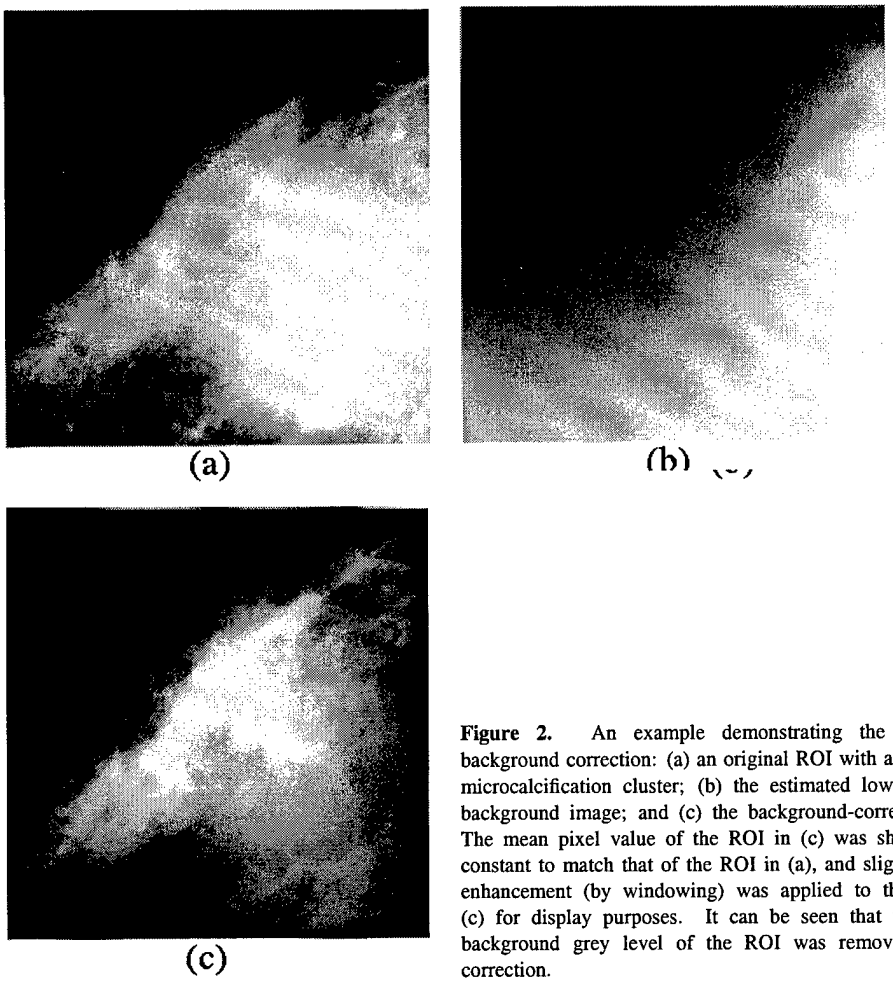


Figure 2. An example demonstrating the effect of background correction: (a) an original ROI with a malignant microcalcification cluster; (b) the estimated low-frequency background image; and (c) the background-corrected ROI. The mean pixel value of the ROI in (c) was shifted by a constant to match that of the ROI in (a), and slight contrast enhancement (by windowing) was applied to the ROI in (c) for display purposes. It can be seen that the sloped background grey level of the ROI was removed by the correction.

2.2. *Texture features*

Our previous studies indicated that the texture features derived from the spatial grey level dependence matrix (SGLD) (Haralick *et al* 1973), also known as the concurrence matrix or the co-occurrence matrix, of the ROI were useful in classification of masses and normal breast tissue (Cheng *et al* 1994, Petrosian *et al* 1994, Chan *et al* 1995b). We further expanded the texture feature space to include multi-distance features and obtained improved results (Wei *et al* 1995a). In this study, we applied texture analysis to the evaluation of textural changes in the breast tissue due to a developing malignancy. The SGLD matrix element, $p_{\theta,d}(i, j)$, is the joint probability of the occurrence of grey levels i and j for pixel pairs which are separated by a distance d and at a direction θ. Because of the discrete

nature of the digital image, the distance d is limited to integral multiples of the pixel size, and the value of θ is limited to 0, 45, 90, and 135° at $d = 1$, and to these and other discrete angles as d increases. We constructed SGLD matrices from pixel pairs in a sub-region of 512×512 pixels centred approximately at the cluster in the background-corrected ROI. Four SGLD matrices, one at each of the four directions, 0, 45, 90, and 135°, were constructed for a given pixel pair distance. The pixel pair distance was varied from four to 40 pixels in increments of four pixels. Therefore, a total of 40 SGLD matrices were derived from each ROI.

The SGLD matrix depends on the bin width (or grey level interval) used in accumulating the histogram. We found in our previous mass classification study (Chan *et al* 1995b) that a bin width of 16 grey levels was a reasonable compromise between grey level resolution and statistical noise. In this study, the ROIs had two times more pixels in width and in height than those in our previous studies, resulting in four times as many pixels in each ROI. Thus, we could use a smaller bin width to obtain approximately the same statistics in the SGLD matrices. Furthermore, our previous study on the digitization requirements of mammograms (Chan *et al* 1994a) indicated that at least nine-bit grey level resolution was required for detection of subtle microcalcifications. We therefore chose a bin width of four grey levels for all SGLD matrices in this study. This is equivalent to reducing the grey level resolution (or bit depth) of the 12-bit image to ten bits by eliminating the two least significant bits.

A number of texture features can be derived from an SGLD matrix (Haralick *et al* 1973, Conners 1979). In our previous studies for mass and normal tissue classification (Chan *et al* 1995b, Wei *et al* 1995a), we evaluated eight texture measures: correlation, entropy, energy (angular second moment), inertia, inverse difference moment, sum average, sum entropy, and difference entropy. In this study, we included five additional texture features: difference average, sum variance, difference variance, information measure of correlation 1, and information measure of correlation 2. The mathematical expressions of these 13 texture features are given in the appendix. These features describe the shape of the SGLD matrix and generally contain information about the image characteristics such as homogeneity, contrast, and the presence of organized structures, as well as the complexity and grey level transitions within the image (Haralick *et al* 1973).

As discussed in our previous study (Chan *et al* 1995b), we did not find a significant dependence of the discriminatory power of the texture features on the direction of the pixel pairs for mammographic textures. However, since the actual distance between the pixel pair in the diagonal direction was a factor of $\sqrt{2}$ of that in the axial direction, we averaged the feature values at the axial directions (0 and 90°) and also at the diagonal directions (45 and 135°) separately for each texture measure derived from the SGLD matrix at a given pixel pair distance. The average texture features at the ten pixel pair distances therefore formed a 260-dimensional feature space for the classification task.

2.3. Feature selection

The dimension of the texture feature space derived from the SGLD matrices at different pixel distances and directions is very large. It is well known that the presence of ineffective features often degrades classifier performance, especially when the training data set is small (Raudys and Pikelis 1980, Fukunaga and Hayes 1989). Investigators in CAD research have employed different methods for feature selection. Goldberg *et al* (1992) selected features for classifying malignant and benign masses on ultrasound images by evaluation of the discriminatory ability of the individual features. Wu *et al* (1993) selected features based

on the difference in the average values of the individual features between the two classes. Lo *et al* (1995) ranked the importance of each feature based on its effect on the classification accuracy, and then eliminated the features, one at a time, from the least important to the most important, to determine the smallest set of features that provided the highest classification accuracy in their data set.

The stepwise procedure in linear discriminant analysis is an established method for selection of useful features for a classification task (Norusis 1993). In our previous studies, we have employed stepwise feature selection and successfully selected a small number of effective features from very large feature spaces (Chan *et al* 1995b, Wei *et al* 1995a). A detailed description of this procedure can be found in the literature. Briefly, one feature is added to or removed from the selected feature set in alternate steps. The effect of the feature on the separation of the two groups is analysed using the Wilks lambda criterion (minimization of the ratio of the within-group sum of squares to the total sum of squares of the two class distributions). The significance of the change in the Wilks lambda when a feature is added to or removed from the model is estimated by F statistics. The user can choose the values of two parameters, the F-to-enter threshold (F_{in}) and the F-to-remove threshold (F_{out}), to control the number of features to be selected. In the feature entry step, each of the features not yet in the model is entered one at a time. The feature variable that causes the most significant change in the Wilks lambda will be included in the feature set if the F value is greater than the F_{in} threshold. In the feature removal step, each of the features already in the model is removed one at a time. The feature variable that causes the least significant change in the Wilks lambda will be excluded from the feature set if the F value is below the F_{out} threshold. The stepwise procedure terminates when the F values for all features not in the model are smaller than the F_{in} threshold and the F values for all features in the model are greater than the F_{out} threshold. Therefore, the number of selected features will decrease if either the F_{in} threshold or the F_{out} threshold is increased. Since the optimal values of the two F thresholds are not known *a priori*, we varied these two thresholds over a wide range to obtain feature sets containing different number of features. The classification accuracies of the different feature sets were then evaluated as described below.

2.4. The artificial neural network (ANN)

We used a feed-forward backpropagation ANN for feature classification in the texture feature space. In this ANN, the nodes are organized in an input layer, an output layer, and one or more hidden layers as shown in figure 3. The nodes are interconnected by weights and information propagates from one layer to the next through a sigmoidal activation function. The learning of the ANN is a supervised process in which known training cases are input to the ANN and the weights are adjusted with an iterative backpropagation procedure in order to achieve a desired input–output relationship. Detailed description of the backpropagation algorithm can be found in the literature (Freeman and Skapura 1991).

To improve the convergence rate and the stability of training, we implemented batch processing in which the weight changes obtained from each training case were accumulated and the weights were updated after the entire set of training cases was evaluated. The batch processing method improves the stability with a tradeoff in the convergence rate. To improve the convergence rate, we included a momentum term and used the delta-bar-delta rule for updating the weights (Sahiner *et al* 1996b). The updated weight is given by

$$w_i(t+1) = w_i(t) - \eta_i(t)\Delta w_i(t)$$

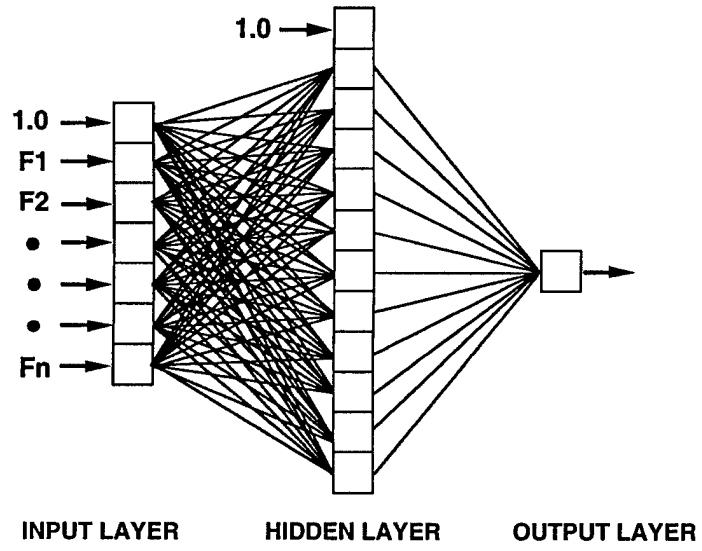


Figure 3. A schematic diagram of the backpropagation neural network classifier used in this study. The number of input nodes was equal to the number of input features. The number of hidden nodes could be varied to obtain the best performance. One output node was used in all ANNs. An ANN with I input nodes, H hidden nodes, and one output node will be denoted as I–H–1.

where $\eta_i(t)$ is the learning rate, $w_i(t)$ is the weight and $\Delta w_i(t)$ is the weight increment for the ith node at training epoch t. When $\eta_i(t)$ is small, the learning is slow but stable. When $\eta_i(t)$ is large, learning is fast but can be unstable. In the delta-bar-delta rule, $\eta_i(t)$ is adjusted adaptively based on the weight increments in two consecutive epochs.

If $\Delta w_i(t-1)\Delta w_i(t) > 0$, $\eta_i(t-1)$ is too small and can be increased:

$$\eta_i(t) = \eta_i(t-1) + \varepsilon \qquad \varepsilon > 0.$$

If $\Delta w_i(t-1)\Delta w_i(t) < 0$, $\eta_i(t-1)$ is too large and should be reduced by a factor r:

$$\eta_i(t) = \eta_i(t-1)r \qquad 0 < r < 1.$$

In this study, we applied a leave-one-out method to training and testing of the ANN classifier. If a data set with N samples is available for training and testing, $(N-1)$ samples will be used for training the classifier and the trained classifier will be evaluated with the left-out test sample. The procedure is repeated N times, each time with a different left-out sample. The test results of the N samples are accumulated to form a distribution of test scores. In the present study, all images of the same patient were left out as test samples in each training cycle and the images from the other $(N-1)$ patients were used for training. The results of all test images from the N training cycles were accumulated to form a distribution of test scores.

Another commonly used method for training and testing a classifier with a small data set is a cross-validation method (Weiss and Kulilowski 1991). In this method, the data set is randomly partitioned into a training set and a test set with a specified training-to-test-case ratio. The training and testing of the classifier are then performed with the partitioned training and test sets, respectively. To reduce the dependence on the training and test cases, the procedure is repeated many times with different partitioning. The results are

then averaged over the many partitions to obtain an estimate of the classifier performance. We performed a limited study using the cross-validation method and compared the results with the leave-one-out method. To ensure independence of the training and test sets in the cross-validation method, the case partitioning was performed with the constraint that images of the same patient were always grouped into the same set.

The performance of the ANN classifier was evaluated by ROC methodology (Swets and Pickett 1982, Metz 1986). The output value of the ANN was used as the decision variable in the ROC analysis. An ROC curve, which is the relationship between the true-positive fraction (TPF) and false-positive fraction (FPF), could be generated by setting different decision thresholds on the output values of the ANN. In this study, we used the LABROC program (Metz *et al* 1990), which assumes binormal distributions of the decision variable for the normal and abnormal cases and fits an ROC curve based on maximum-likelihood estimation, to estimate the area under the ROC curve (A_z) and the standard deviation (SD) of A_z. A_z was used as an index of classification accuracy. For the leave-one-out method, the test A_z was obtained from analysis of the accumulated test score distribution from all N cycles. For the cross-validation method, the average performance of the ANN was estimated as the average of the 50 test A_z values obtained from training and testing with 50 different partitions of the data sets.

3. Results

Some representative subsets of features selected by the stepwise procedure from the 260-dimensional texture feature space are listed in table 1. The number of features was varied by changing the F_{in} and F_{out} thresholds as shown in the table. The number of features selected usually remained constant over a range of F_{in} and F_{out} thresholds. For example, there were six selected features when (F_{in}, F_{out}) were reduced from about (2.65, 2.55) to (2.1, 2.0), and seven selected features when reduced from about (1.9, 1.8) to (0.56, 0.55). When the F_{in} and F_{out} were reduced slightly further, the number of selected features increased abruptly to 19.

We evaluated each feature subset by using the feature subset as input to the ANN and estimating the classification accuracy A_z. For a given feature set containing I features, an ANN with I input nodes, one to ten hidden nodes, and one output node was trained with the leave-one-case-out method as described above. For each training cycle with $(N-1)$ training cases, the ANN was trained up to 30 000 epochs. The test result for the left-out case was obtained at fixed intervals of epochs (e.g., every 1000 epochs). After the N training cycles were completed, the test results of the entire dataset would have been accumulated at the fixed intervals of epochs. Therefore, an ROC curve could be fitted to the output of the test cases and the A_z estimated at the fixed intervals of epochs. Figure 4 shows the typical convergence trend of the test A_z results as the training epochs increased. The test A_z generally increased rapidly for the first few thousand epochs and then levelled off gradually. In this example, the test A_z remained at a constant level of about 0.88 when the ANN was trained for more than 8000 epochs. In some cases, the test A_z decreased if the ANN was over-trained. The test A_z values reported in the following discussion were obtained at the maximum plateau region.

The dependence of the classification accuracy, A_z, on the ANN architecture is shown in figure 5 for the different feature subsets. For convenience of comparison, an ANN without a hidden layer was plotted as an ANN with zero hidden nodes. The number of hidden nodes for a three-layer ANN was varied from one to ten. The standard deviation (SD) of the A_z, estimated by the LABROC1 program, ranged from 0.035 to 0.045. For a given feature set,

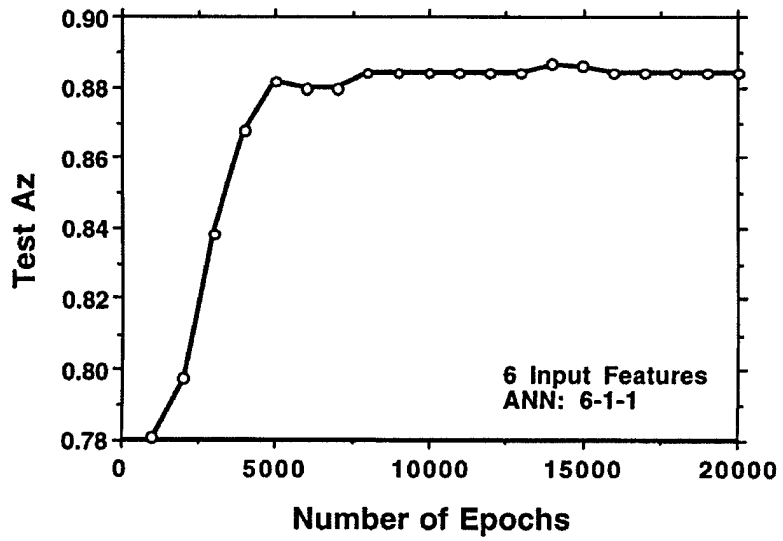


Figure 4. An example demonstrating the dependence of test A_z on the number of training epochs. The test A_z generally increased rapidly during the first 5000 epochs and then gradually reached a plateau or a broad maximum.

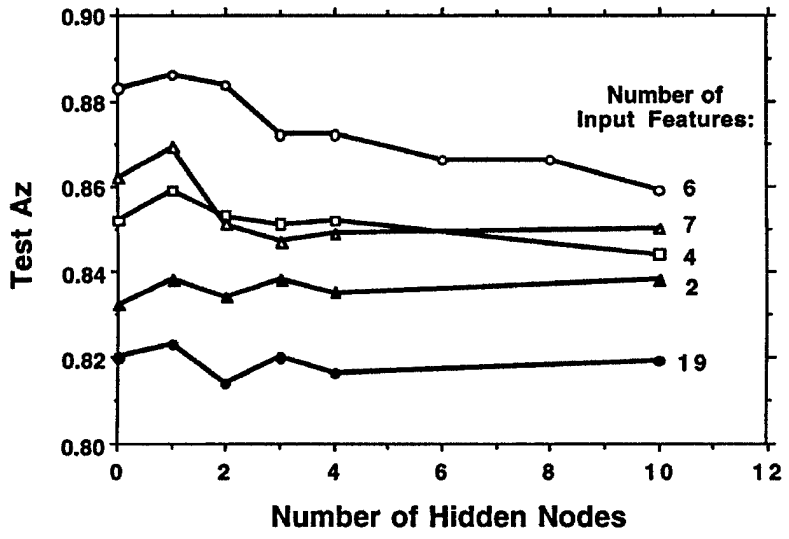


Figure 5. The dependence of the classification accuracy, A_z, on the number of hidden nodes in the ANN classifier. To facilitate comparison, the results for a two-layer ANN that had no hidden layer were plotted as data points with zero hidden nodes. The ANNs with one hidden node consistently provided higher accuracy than the other ANNs for all input feature sets. The input feature set with six selected features was the most effective in classifying malignant and benign microcalcifications among the selected feature sets.

Table 1. Texture features selected by stepwise feature selection procedure for different F-to-enter (F_{in}) and F-to-remove (F_{out}) thresholds.

$F_{in} = 3.84$ $F_{out} = 2.71$	$F_{in} = 2.7$ $F_{out} = 2.6$	$F_{in} = 2.5$ $F_{out} = 2.4$	$F_{in} = 1.7$ $F_{out} = 1.5$	$F_{in} = 0.55$ $F_{out} = 0.45$
Diff. entropy ($d = 8$)	Correlation ($d = 40$, diagonal)	Diff. average ($d = 4$)	Correlation ($d = 40$, diagonal)	Correlation ($d = 8$)
Inv. diff. moment ($d = 4$)	Diff. entropy ($d = 8$)	Diff. entropy ($d = 8$)	Diff. average ($d = 4$)	Diff. average ($d = 32$)
	Inertia ($d = 40$)	Diff. entropy ($d = 32$, diagonal)	Diff. entropy ($d = 8$)	Diff. average ($d = 4$)
	Inv. diff. moment ($d = 4$)	Inertia ($d = 4$)	Diff. entropy ($d = 32$, diagonal)	Diff. average ($d = 40$, diagonal)
		Inertia ($d = 40$)	Inertia ($d = 40$)	Diff. entropy ($d = 8$)
		Inv. diff. moment ($d = 12$)	Inv. diff. moment ($d = 12$)	Diff. entropy ($d = 32$, diagonal)
			Inv. diff. moment ($d = 4$)	Diff. variance ($d = 40$, diagonal)
				Energy ($d = 24$, diagonal)
				Information measure of correlation 1 ($d = 36$)
				Information measure of correlation 1 ($d = 40$, diagonal)
				Information measure of correlation 2 ($d = 24$)
				Information measure of correlation 2 ($d = 36$)
				Information measure of correlation 2 ($d = 4$, diagonal)
				Inertia ($d = 4$)
				Inertia ($d = 40$)
				Inv. diff. moment ($d = 12$)
				Inv. diff. moment ($d = 8$)
				Inv. diff. moment ($d = 4$, diagonal)
				Inv. diff. moment ($d = 8$, diagonal)

the variation of the A_z values with the number of ANN hidden nodes was within one SD. However, the maximum A_z consistently occurred at the ANN with one hidden node for all feature sets. The feature set with six features provided the highest A_z over the entire range of hidden nodes studied. The maximum A_z of 0.88 was obtained with an ANN of six input nodes, one hidden node, and one output node. The ROC curves that had the two highest A_z values obtained with six and seven input features and one hidden node are plotted in figure 6.

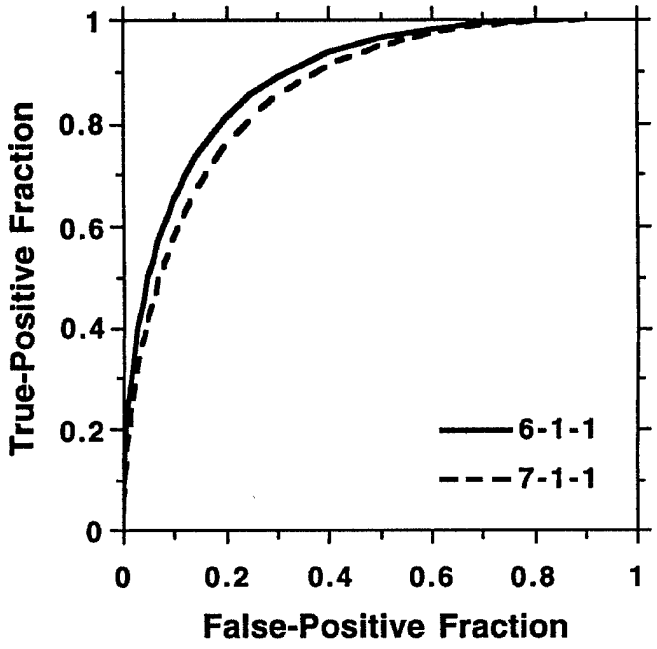


Figure 6. ROC curves that had the two highest A_z values obtained with the six-feature and seven-feature sets and one hidden node shown in figure 5.

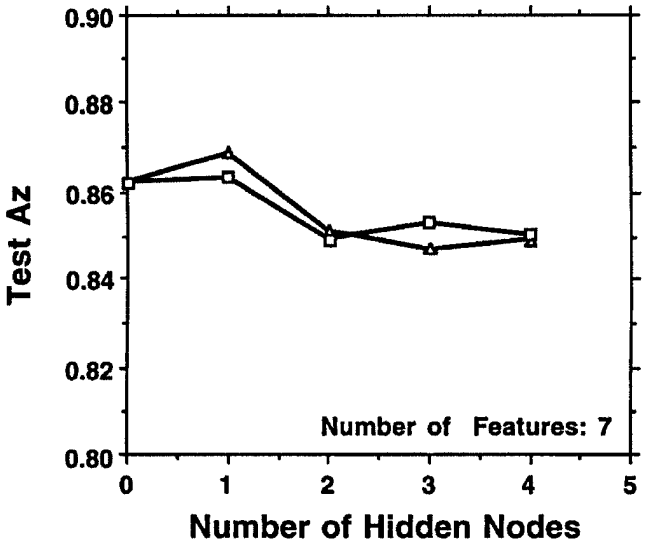


Figure 7. The effect of the initialization of the weights in the ANN on classifier performance. An ANN with seven input nodes, zero to four hidden nodes, and one output node was studied. The two data points at each ANN configuration represent the two different initializations of its weights. The difference in the initial weights appears to have very small effect on the convergence of the ANN.

To evaluate the variation of the classification accuracy on the initialization of the ANN, we used two different random number seeds to generate the initial weights for the ANNs with seven input features. The A_z values are plotted in figure 7 for the ANNs with different numbers of hidden nodes. The differences in A_z were within 0.01 for the different ANNs, indicating that the initial weights do not have a strong effect on the convergence of the ANNs.

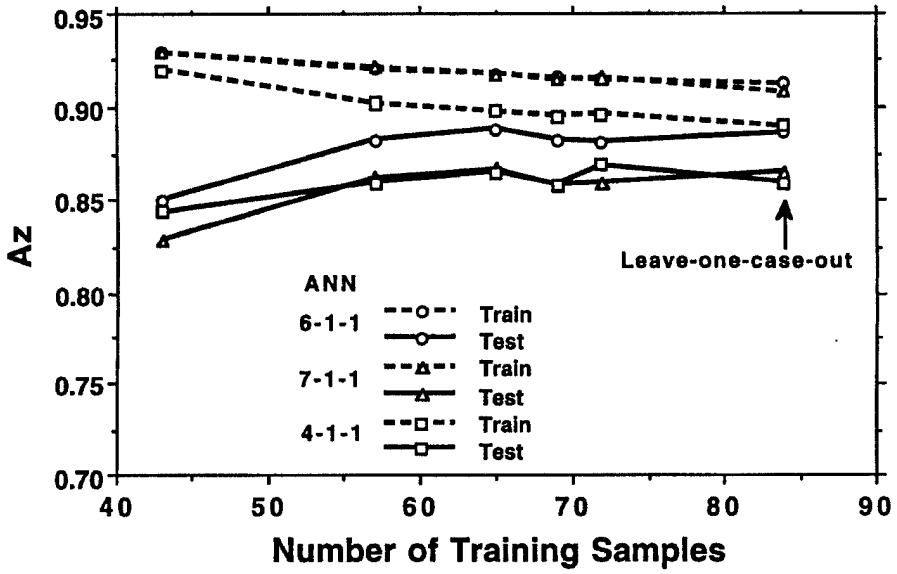


Figure 8. The dependence of A_z on the number of training cases obtained from a cross-validation method. The number of training cases was varied by randomly partitioning the data set into a training set and a test set with training-to-test-sample ratios of one to five. For a given training-to-test-sample ratio, the training (or test) A_z plotted was the average of the 50 A_z values obtained from the 50 random partitions of the data set. For comparison, the A_z values obtained with the leave-one-case-out training and test method were also plotted as the data points with 84 training samples.

Figure 8 shows the performance of the ANN classifiers which were trained and tested with a cross-validation method. The number of input nodes of the ANNs corresponded to the number of input features; the numbers of hidden nodes and output nodes were both set to be one. The training-to-test sample ratio was varied from one to five. The data set was randomly partitioned 50 times at each ratio and the mean training and test A_z values from the 50 partitions were plotted against the number of training samples. Because of the constraint that films of the same patient were always grouped into the same set, the number of training (or test) samples in each of the 50 partitions might not be equal: the expected number of training samples calculated as the nearest integer of $[86R/(R+1)]$, where R is the training-to-test-sample ratio, was plotted as the abscissae. As the training-to-test-sample ratios increased from one to five, the expected number of training samples increased from 43 to 72. To facilitate comparison, the A_z for the corresponding ANN classifiers trained with the leave-one-case-out method was plotted as the data point having an expected number of training samples of 84.

It can be seen that the training A_z decreased slowly as the number of training samples increased. The test A_z, on the other hand, increased as the number of training samples

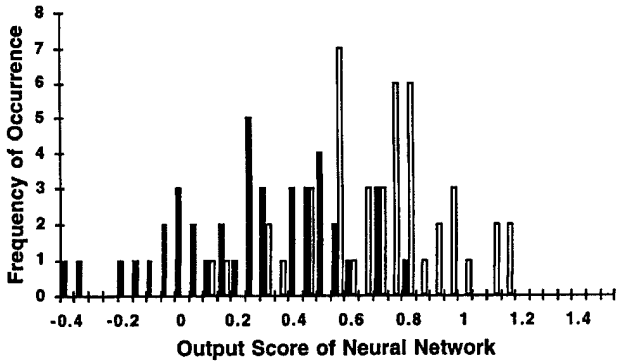
increased. Because the number of test samples was small when the ratio was large, the SD for each test A_z ranged from 0.06 to 0.12 when the ratio increased from one to five. However, the SDs of the mean test A_z values from the 50 partitions varied from 0.01 to 0.02. It can be seen that the fluctuations of the data points were within one SD of the mean. The trend of the curves generally agrees with the expectations that small training sets over-estimate the classifier performance and the trained classifiers perform poorly on test sets, and that both the training and test results will approach the 'true' performance as the number of training samples approaches infinity (Raudys and Pikelis 1980, Fukunaga and Hayes 1989).

The output scores of the ANN with six input features, one hidden node, and one output node for the 86 test samples obtained with the leave-one-case-out method are plotted in figure 9(a). The output scores of the ANN have been scaled linearly for the purpose of plotting the graph. The linear transformation simply expands the horizontal scale without any effect on the relative distribution of the scores. It can be seen that there was good separation between the malignant and benign clusters. If the decision threshold was set at 0.85, 11 of the 45 benign samples were correctly classified without any false negatives (a sensitivity of 100% at a specificity of 24%). At a decision threshold of 0.75, 23 of the benign samples were correctly classified but one malignant sample was missed (a sensitivity of 98% at a specificity of 51%). When the ANN output scores were analysed with the LABROC1 program, the area under the fitted ROC curve was 0.88.

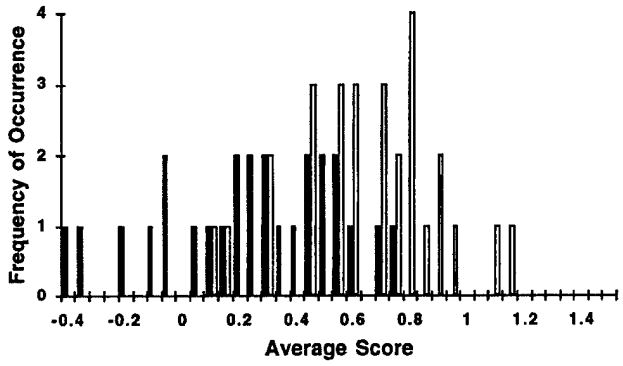
Because some of the samples are films from the same patient, it will be reasonable to make the malignant or benign decision on a case-by-case basis. Two approaches were investigated: one used the average score from all films of the same patient and the other used the minimum score from all films of the same patient for decision making. The latter was a more conservative approach because a lower score corresponded to higher likelihood of malignancy in our analysis. The distributions of the average scores and the minimum scores for the 54 cases are shown in figure 9(b) and (c), respectively. If a decision threshold were set at an average score of 0.80, ten of the 28 benign cases would be correctly classified without any false negative (a sensitivity of 100% at a specificity of 36%). Alternatively, if a decision threshold was set at a minimum score of 0.75, 11 of the 28 benign cases would be correctly classified without missing any malignant cases (a sensitivity of 100% at a specificity of 39%).

4. Discussion

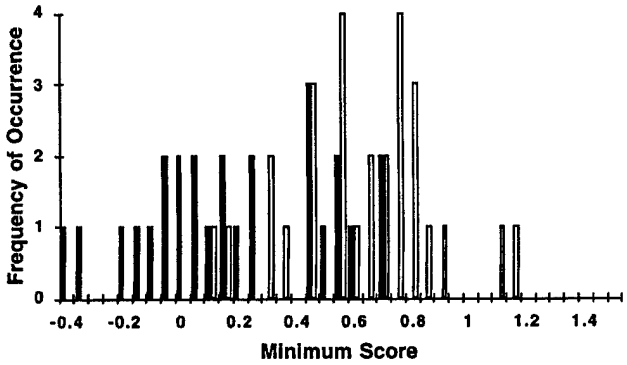
We have investigated the usefulness of texture analysis in predicting the malignant and benign nature of abnormal breast tissue containing clustered microcalcifications. All case samples used in this study had been surgically biopsied, indicating that definitive diagnosis could not be made by the mammographic appearance of the benign clusters. Our results show that there are changes in the texture of the breast tissue in which a malignancy is developing, and that these changes can be distinguished from the benign tissue texture by computerized analysis although their differences are not visually apparent on mammograms. Based on the results of texture analysis and ANN classification, a significant fraction of benign cases can be correctly identified. This information may be used to reduce the number of biopsies, thereby improving the positive predictive value of mammography. Our preliminary study therefore demonstrates that computerized classification may be a useful aid in mammographic interpretation. Further investigation to determine if this approach can be generalized to large data sets is warranted.



(a)



(b)



(c)

Figure 9. The distributions of the discriminant scores for the malignant (black) and benign (white) microcalcification clusters. The test results for the ANN with six input features, one hidden node, and one output node trained with the leave-one-case-out method are shown. The output scores of the ANN have been scaled linearly for the purpose of plotting. (a) Distribution of the output scores from the ANN classifier for 86 test samples; (b) distribution of the average scores for 54 cases; (c) distribution of the minimum scores for 54 cases.

We used an ANN as a feature classifier for this classification task. By varying the structure of the ANN, both linear and non-linear classifiers could be studied. An analysis of the dependence of the classification accuracy on ANN architecture (figure 5) indicated that ANNs with one hidden node provided the best performance for all feature sets. Because an ANN with one hidden node is equivalent to a linear classifier, the results appear to indicate that a linear classifier may be the optimal choice for this classification task. However, it should be cautioned that the performance of a classifier depends on the number of training samples relative to the number of parameters to be trained in the classifier (Raudys and Pikelis 1980, Fukunaga and Hayes 1989). Since the data set in this study was small and the number of weights to be trained in an ANN increased rapidly with the number of hidden nodes, the observed reduction in classification accuracy with the increasing number of hidden nodes could be caused by insufficient training samples. The optimal choice of a feature classifier for this classification task will have to be investigated further when a large data set is available.

Thiele *et al* (1996) recently studied the classification of the tissue texture surrounding calcification clusters to predict malignant or benign outcomes. They used texture measures calculated from the SGLD matrices and fractal geometry as input to a linear discriminant classifier or a logistic discriminant classifier. Their results also demonstrated that texture analysis showed significant discriminatory power between benign and malignant tissue. In a data set of 54 cases (36 benign, 18 malignant), they obtained a sensitivity of 89% at a specificity of 83%. In their calculation of the SGLD matrices in the tissue region, they included subtle microcalcifications but excluded the pixels containing large and bright calcifications by manually identifying the calcification areas with grey level thresholding. In our SGLD matrix calculation, all pixels in the 512×512 ROI containing the microcalcification cluster were included. Because of the many differences between the two studies and the difference in the data set, it is not known which approach will provide more effective texture features. However, the advantage of our approach is that no manual identification of individual microcalcifications is needed and the analysis can be much more efficient. Minimal operator intervention will be a practical consideration if the computerized classification technique is to be implemented in clinical settings.

In this study, we performed background correction in a 1024×1024 ROI but calculated texture features in a subregion of 512×512 pixels centred approximately at the cluster of microcalcifications. The use of a subregion smaller than the original 1024×1024 ROI would avoid any potential edge effects caused by background correction. Furthermore, because many of the clusters in our data set could be enclosed by a 512×512 region, calculation of texture features in the original ROI would average the texture features in the cluster region with those in a large region of possibly normal tissue. The choice of the subregion size was subjective in this study, taking into consideration the tradeoff between the averaging effect and the statistics needed in the SGLD matrix formation. Whether a different choice of the region size, or use of variable size according to the cluster diameter, would improve the effectiveness of the texture features remains to be studied.

In this study, we did not perform a systematic optimization of the parameters for texture extraction. Many of the parameters were chosen based on our experience in other applications. The goal of this study is to demonstrate the feasibility of using computerized texture analysis for classification of malignant and benign microcalcifications. Our results indicate that the SGLD texture features are useful in such an application although the techniques have not been optimized. In future studies, both the feature extraction techniques and the classifier should be improved by optimization of the various parameters using a large data set.

5. Conclusion

We have developed a computerized method for classification of malignant and benign microcalcification clusters on mammograms. The computer extracts texture features from an ROI containing the microcalcification cluster and predicts its pathology using a trained neural network classifier. The effectiveness of our approach has been demonstrated with a small data set. The classifier could correctly identify a significant fraction of benign cases, which had been recommended for surgical biopsy under current clinical criteria, without missing any malignant cases. The computerized texture analysis may therefore provide useful information for reducing the number of negative biopsies. Further investigation will be conducted with a larger data set to determine the generalizability of these results. The combination of this texture classification method with other morphological features or patient information will be investigated. The optimization of the classifier design will also be examined.

Acknowledgments

This work is supported by USPHS grant CA 48129. The authors are grateful to Datong Wei, PhD, for his assistance in the initial stage of this study, and Charles E Metz, PhD, for the LABROC1 program.

Appendix. The spatial grey level dependence (SGLD) matrix and texture features

The (i, j)th element of the SGLD matrix, $p_{\theta,d}(i, j)$, is the joint probability that the grey levels i and j occur in a direction of angle θ and at a distance of d pixels apart over the entire ROI. The joint probability $p_{\theta,d}(i, j)$ is normalized by the number of grey level pairs obtained from the ROI with a pixel distance of d. For each ROI, thirteen texture measures were derived from its SGLD matrix as described below. Most of the expressions can be found in the literature (Haralick *et al* 1973). Some differences in the expressions may be noted. A simplified notation $p(i, j)$ will be used to denote the SGLD matrix elements in the following equations.

$$\text{Energy} = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}[p(i, j)]^2 \tag{A1}$$

where n is the number of grey levels in the image.

$$\text{Correlation} = \left(\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i - \mu_x)(j - \mu_y)p(i, j)\right) \Big/ (\sigma_x\sigma_y) \tag{A2}$$

where

$$\mu_x = \sum_{i=0}^{n-1} i p_x(i) \qquad \sigma_x^2 = \sum_{i=0}^{n-1}(i - \mu_x)^2 p_x(i)$$

$$\mu_y = \sum_{j=0}^{n-1} j p_y(j) \qquad \sigma_y^2 = \sum_{j=0}^{n-1}(j - \mu_y)^2 p_y(j)$$

are the mean and variance of the marginal distributions $p_x(i)$ and $p_y(j)$, respectively.

$$p_x(i) = \sum_{j=0}^{n-1} p(i, j)$$

$$p_y(j) = \sum_{i=0}^{n-1} p(i, j).$$

$$\text{Inertia} = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(i-j)^2 p(i, j) \tag{A3}$$

$$\text{Entropy} = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p(i, j)\log_2 p(i, j) \tag{A4}$$

$$\text{Inverse difference moment} = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\frac{1}{1+(i-j)^2}p(i, j) \tag{A5}$$

$$\text{Sum average} = \sum_{k=0}^{2n-2} k p_{x+y}(k) \tag{A6}$$

where

$$p_{x+y}(k) = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p(i, j) \qquad i+j=k \qquad k=0, \ldots, 2n-2.$$

$$\text{Sum variance} = \sum_{k=0}^{2n-2}(k - \text{sum average})^2 p_{x+y}(k) \tag{A7}$$

$$\text{Sum entropy} = -\sum_{k=0}^{2n-2} p_{x+y}(k)\log_2 p_{x+y}(k) \tag{A8}$$

$$\text{Difference average} = \sum_{k=0}^{n-1} k p_{x-y}(k) \tag{A9}$$

where

$$p_{x-y}(k) = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p(i, j) \qquad |i-j|=k \qquad k=0, \ldots, n-1.$$

$$\text{Difference variance} = \sum_{k=0}^{n-1}(k - \text{difference average})^2 p_{x-y}(k) \tag{A10}$$

$$\text{Difference entropy} = -\sum_{k=0}^{n-1} p_{x-y}(k)\log_2 p_{x-y}(k) \tag{A11}$$

$$\text{Information measure of correlation 1} = (\text{entropy} - H_1)/\max\{H_x, H_y\} \tag{A12}$$

where

$$H_1 = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p(i, j)\log_2[p_x(i)p_y(j)]$$

$$H_x = -\sum_{i=0}^{n-1} p_x(i)\log_2 p_x(i)$$

$$H_y = -\sum_{j=0}^{n-1} p_y(j) \log_2 p_y(j).$$

$$\text{Information measure of correlation 2} = \sqrt{1 - \exp[-2(H_2 - \text{entropy})]} \tag{A13}$$

where

$$H_2 = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_x(i) p_y(j) \log_2[p_x(i) p_y(j)].$$

References

Ackerman L V and Gose E E 1972 Breast lesion classification by computer and xeroradiograph *Cancer* **30** 1025–35

Ackerman L V, Mucciardi A N, Gose E E and Alcorn F S 1973 Classification of benign and malignant breast tumors on the basis of 36 radiographic properties *Cancer* **31** 342–52

Adler D D and Helvie M A 1992 Mammographic biopsy recommendations *Current Opinion Radiol.* **4** 123–9

Baker J A, Kornguth P J, Lo J Y and Floyd C E 1996 Artificial neural network: improving the quality of breast biopsy recommendations *Radiology* **198** 131–5

Chan H-P, Niklason L T, Ikeda D M and Adler D D 1992 Computer-aided diagnosis in mammography: detection and characterization of microcalcifications *Med. Phys.* **19** 831

Chan H-P, Niklason L T, Ikeda D M, Lam K L and Adler D D 1994a Digitization requirements in mammography: effects on computer-aided detection of microcalcifications *Med. Phys.* **21** 1203–11

Chan H-P, Sahiner B, Lam K L, Wei D, Helvie M A and Adler D D 1995a Classification of malignant and benign microcalcifications on mammograms using an artificial neural network *Proc. World Congress on Neural Networks (Washington, DC, 1995)* vol 2 (Mahwah, NJ: INNS) pp 889–92

Chan H-P, Wei D, Helvie M A, Sahiner B, Adler D D, Goodsitt M M and Petrick N 1995b Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space *Phys. Med. Biol.* **40** 857–76

Chan H-P, Wei D, Lam K L, Lo S-C B, Sahiner B, Helvie M A and Adler D D 1995c Computerized detection and classification of microcalcifications on mammograms *Proc. SPIE* **2434** 612–20

Chan H P, Wei D, Lam K L, Sahiner B, Helvie M A, Adler D D and Goodsitt M M 1995d Classification of malignant and benign microcalcifications by texture analysis *Med. Phys.* **22** 938

Chan H-P, Wei D, Niklason L T, Helvie M A, Lam K L, Goodsitt M M and Adler D D 1994b Computer-aided classification of malignant/benign microcalcifications in mammography *Med. Phys.* **21** 875

Cheng S N C, Chan H P, Helvie M A, Goodsitt M M, Adler D D and St Clair D 1994 Classification of mass and non-mass regions on mammograms using artificial neural network *J. Imaging Sci. Technol.* **38** 598–603

Chitre Y, Dhawan A P and Moskowitz M 1993 Artificial neural network based classification of mammographic microcalcifications using image structure features *Int. J. Pattern Recognition Artificial Intell.* **7** 1377–401

Conners R W 1979 Towards a set of statistical features which measure visually perceivable qualities of textures *Proc. IEEE Conf. on Pattern Recognition and Image Processing* (New York: IEEE) pp 382–90

D'Orsi C J, Getty D J, Swets J A, Pickett R M, Seltzer S E and McNeil B J 1992 Reading and decision aids for improved accuracy and standardization of mammographic diagnosis *Radiology* **184** 619–22

Fox S H, Pujare U M, Wee W G, Moskowitz M and Hutter R V P 1980 A computer analysis of mammographic microcalcifications: global approach *Proc. IEEE 5th Int. Conf. on Pattern Recognition* (New York: IEEE) pp 624–31

Freeman J A and Skapura D M 1991 *Neural Networks—Algorithms, Applications, and Programming Techniques* (Reading, MA: Addison-Wesley)

Fukunaga K and Hayes R R 1989 Effects of sample size on classifier design *IEEE Trans. Pattern Anal. Machine Intell.* **11** 873–85

Gale A G, Roebuck E J, Riley P and Worthington B S 1987 Computer aids to mammographic diagnosis *Br. J. Radiol.* **60** 887–91

Getty D J, Pickett R M, D'Orsi C J and Swets J A 1988 Enhanced interpretation of diagnostic images *Invest. Radiol.* **23** 240–52

Goldberg V, Manduca A, Ewert D L, Gisvold J J and Greenleaf J F 1992 Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence *Med. Phys.* **19** 1475–81

Haralick R M, Shanmugam K and Dinstein I 1973 Texture features for image classification *IEEE Trans. Syst. Man Cybernet.* **3** 610–21

Huo Z, Giger M L, Vyborny C J, Bick U, Lu P, Wolverton D E and Schmidt R A 1995 Analysis of spiculation in the computerized classification of mammographic masses *Med. Phys.* **22** 1569–79

Jiang Y, Nishikawa R M, Wolverton D E, Metz C E, Giger M L, Schmidt R A, Vyborny C J and Doi K 1996 Malignant and benign clustered microcalcifications: automated feature analysis and classification *Radiology* **198** 671–78

Kilday J, Palmieri F and Fox M D 1993 Classifying mammographic lesions using computerized image analysis *IEEE Trans. Med. Imaging* **12** 664–9

Kopans D B 1991 The positive predictive value of mammography *Am. J. Roentgenology* **158** 521–6

Lo J Y, Baker J A, Kornguth P J and Floyd C E 1995 Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features *Acad. Radiol.* **2** 841–50

Metz C E 1986 ROC methodology in radiologic imaging *Invest. Radiol.* **21** 720–33

Metz C E, Shen J H and Herman B A 1990 New methods for estimating a binormal ROC curve from continuously-distributed test results *Annu. Meeting Am. Stat. Assoc. (Anaheim, CA, 1990)*

Norusis M J 1993 *SPSS for Windows Professional Statistics* release 6.0. (Chicago, IL: SPSS)

Petrosian A, Chan H P, Helvie M A, Goodsitt M M and Adler D D 1994 Computer-aided diagnosis in mammography: classification of masses and normal tissue by texture analysis *Phys. Med. Biol.* **39** 2273–88

Raudys S and Pikelis V 1980 On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition *IEEE Trans. Pattern Anal. Machine Intell.* **2** 242–52

Sahiner B, Chan H P, Petrick N, Helvie M A, Adler D D and Goodsitt M M 1996a Classification of masses on mammograms using rubber-band straightening transform and feature analysis *Proc. SPIE* **2710** 44–50

Sahiner B, Chan H P, Petrick N, Wei D, Helvie M A, Adler D D and Goodsitt M M 1996b Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images *IEEE Trans. Med. Imaging* **15** 598–610

Shen L, Rangayyan R M and Desautels J E L 1994 Application of shape analysis to mammographic calcifications *IEEE Trans. Med. Imaging* **13** 263–74

Swets J A and Pickett R M 1982 *Evaluation of Diagnostic System: Methods from Signal Detection Theory* (New York: Academic)

Thiele D L, Kimme-Smith C, Johnson T D, McCombs M and Bassett L W 1996 Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes *Med. Phys.* **23** 549–55

Wee W G, Moskowitz M, Chang N-C, Ting Y-C and Pemmeraju S 1975 Evaluation of mammographic calcifications using a computer program. *Radiology* **116** 717–20

Wei D, Chan H P, Helvie M A, Sahiner B, Petrick N, Adler D D and Goodsitt M M 1995a Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis *Med. Phys.* **22** 1501–13

——1995b Multiresolution texture analysis for classification of mass and normal breast tissue on digital mammograms *Proc. SPIE* **2434** 606–11

Weiss S M and Kulilowski C A 1991 *Computer Systems that Learn* (San Mateo, CA: Morgan Kaufmann)

Wu Y, Freedman M T, Hasegawa A, Zuurbier R A, Lo S C B and Mun S K 1995 Classification of microcalcifications in radiographs of pathologic specimens for the diagnosis of breast cancer *Acad. Radiol.* **2** 199–204

Wu Y, Giger M L, Doi K, Vyborny C J, Schmidt R A and Metz C E 1993 Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer *Radiology* **187** 81–7

Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis

Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Mark A. Helvie, and Mitchell M. Goodsitt
The University of Michigan, Department of Radiology, Ann Arbor, Michigan 48109-0030

(Received 11 October 1996; accepted for publication 27 January 1998)

A new rubber band straightening transform (RBST) is introduced for characterization of mammographic masses as malignant or benign. The RBST transforms a band of pixels surrounding a segmented mass onto the Cartesian plane (the RBST image). The border of a mammographic mass appears approximately as a horizontal line, and possible spiculations resemble vertical lines in the RBST image. In this study, the effectiveness of a set of directional texture features extracted from the RBST images was compared to the effectiveness of the same features extracted from the images before the RBST. A database of 168 mammograms containing biopsy-proven malignant and benign breast masses was digitized at a pixel size of 100 μm$\times$100 μm. Regions of interest (ROIs) containing the biopsied mass were extracted from each mammogram by an experienced radiologist. A clustering algorithm was employed for automated segmentation of each ROI into a mass object and background tissue. Texture features extracted from spatial gray-level dependence matrices and run-length statistics matrices were evaluated for three different regions and representations: (i) the entire ROI; (ii) a band of pixels surrounding the segmented mass object in the ROI; and (iii) the RBST image. Linear discriminant analysis was used for classification, and receiver operating characteristic (ROC) analysis was used to evaluate the classification accuracy. Using the ROC curves as the performance measure, features extracted from the RBST images were found to be significantly more effective than those extracted from the original images. Features extracted from the RBST images yielded an area (A_z) of 0.94 under the ROC curve for classification of mammographic masses as malignant and benign. [S0094-2405(98)00904-3]

Key words: mammography, computer-aided diagnosis, masses, classification, texture analysis, discriminant analysis, ROC analysis

I. INTRODUCTION

Mammography is the most effective method for detection of early breast cancer.[1] However, the positive predictive value of mammographic diagnosis is only about 15%–30%.[2–5] Biopsies performed for mammographically suspicious nonpalpable breast masses had positive predictive values of 29%,[6] 29%,[7] and 21%[8] in three studies. As the number of patients who undergo mammography increases, it will be increasingly important to improve the positive predictive value of the procedure in order to reduce costs and patient discomfort. A computerized algorithm that can assist radiologists in classification of mammographic abnormalities may reduce benign biopsies.

Masses are important indicators of malignancy on mammograms. In recent years, considerable effort has been devoted to the development of computerized methods for detection and classification of mammographic masses.[5,9–23] Methods for classification of mammographic masses can be categorized into two groups: one based on features extracted by a radiologist,[5,16–18] and the other based on computer-extracted features.[19–23]

Classification methods based on features extracted by a radiologist are usually designed to include all mammographic signs such as masses and microcalcifications. Although mammographic features are essential components of these methods, age[5] and the personal and family history of the patient[16] are also sometimes used. Getty *et al.*[17] designed a classifier based on discriminant analysis and 12 mammographic features extracted by radiologists, and showed that the classifier can substantially increase the radiologist's diagnostic accuracy. Wu *et al.*[18] designed a neural network classifier based on 14 mammographic features extracted by an experienced radiologist, and showed that its performance in classifying benign and malignant lesions was higher than the average performance of attending and resident radiologists. Recently, Baker *et al.*[16] reported the development of a classifier based on BI-RADS features of the American College of Radiology and the personal and family history of the patient. The specificity of their neural network classifier was shown to be significantly higher than that of the radiologists at high sensitivity levels.[16]

Mass classification methods based on computer-extracted features have the advantages of objectivity and consistency, since they rely on computerized methods for the entire analysis. However, they may also be more difficult to design. These methods usually first extract the lesion shape using interactive or automatic methods, and then extract features from the shape and gray-level characteristics of the lesion,

and the surrounding tissue. Brzakovic *et al.*[19] classified computer-detected suspicious regions into one of three categories, namely, benign tumor, malignant tumor, or nontumor, using their shape and intensity variations. Kilday *et al.*[20] extracted mass shapes using interactive gray-level thresholding, and classified them into cancer, cyst, or fibroadenoma categories using shape features and patient age. Pohlman *et al.*[22] used a region growing algorithm for tumor segmentation, and morphological features extracted from the segmented masses for classification. Huo *et al.*[23] developed a technique to quantify the degree of spiculation of a lesion and classified masses as malignant or benign using the spiculation measures. Their computer-extracted spiculation measure was shown to yield higher classification accuracy than the spiculation rating of an experienced radiologist.[23]

Typical characteristics of malignant masses include high density, spiculated margins, and indistinct, irregular or fuzzy contours. Benign breast masses tend to have sharper, well-circumscribed borders.[24] Automatic characterization of the region surrounding a mass is therefore very important in computer aided diagnosis. An important factor in analyzing the gray-level, gradient, spiculation, and texture characteristics of the area around a mass is their directional dependence. For some of these characteristics, it is difficult to preserve significant directional information from the region surrounding the mass. For example, the gradient of the opacity caused by a mass is radially oriented, and this makes it difficult to extract gradient and texture features from the region surrounding the mass without some preprocessing. Similarly, detection of spiculations is complicated by the fact that the search direction for the spiculation changes with the shape of the mass and the curvature of its margin. To overcome this problem, we have designed a rubber band straightening transform (RBST) which maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the border of a mass is expected to appear approximately as a horizontal edge, and spiculations are expected to appear approximately as vertical lines.

The classification algorithm in this paper consisted of four main steps, which were (1) automatic extraction of the mass shape; (2) computation of the RBST image; (3) extraction of texture features; and (4) classification using linear discriminant analysis (LDA). To study the potential advantage of texture feature extraction using the RBST images, the effectiveness of texture features extracted from the RBST images for classification was compared to the effectiveness of the same features extracted from the region surrounding the mass in the original image.

The rest of the paper is organized as follows. In the next section, we describe our image database, and the four steps of the classification algorithm specified above. In Sec. III, we present the classification results using texture features extracted from different image representations (with or without the RBST). Section IV contains a discussion of these results. Finally, Sec. V concludes the investigation and provides a scope for further research.

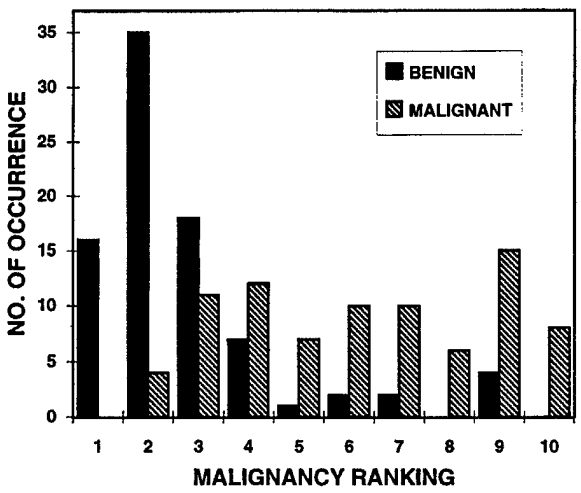


FIG. 1. The distribution of the malignancy ranking of the masses in our dataset, by an experienced radiologist. 1: Very likely benign, 10: Very likely malignant.

II. MATERIALS AND METHODS

A. Data set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsy in the Department of Radiology at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. To avoid the effect of repetitive grid lines on the image texture, mammograms that contained grid lines caused by the stationary grid of some older mammographic units were excluded. The mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm$\times$100 μm, and 4096 gray levels. The digitizer was calibrated so that gray-level values were linearly proportional to the optical density (OD) within the range of 0.1–2.8 o.d. units, with a slope of 0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The o.d. range of the digitizer was 0 to 3.5.

The data set in this study included 168 mammograms from 72 patients. Of the 168 mammograms, 83 contained malignant masses, and 85 contained benign masses. Six of the benign masses and 45 of the malignant masses were spiculated, as determined visually by a radiologist experienced in mammographic interpretation. Regions of interest (ROIs) containing the biopsied masses were extracted by the same radiologist from each mammogram. The size of each ROI was 256$\times$256 pixels. Our data set contained a range of obvious to subtle masses. The probability of malignancy of each mass, based on its mammographic appearance, was ranked by the radiologist on a scale of 1 to 10, where a ranking of 1 corresponded to the masses with the most benign mammographic appearance. The distribution of the malignancy ranking of the masses is shown in Fig. 1.

B. Mass shape extraction

We used a pixel-by-pixel clustering algorithm followed by object selection for segmentation of the ROI into a mass

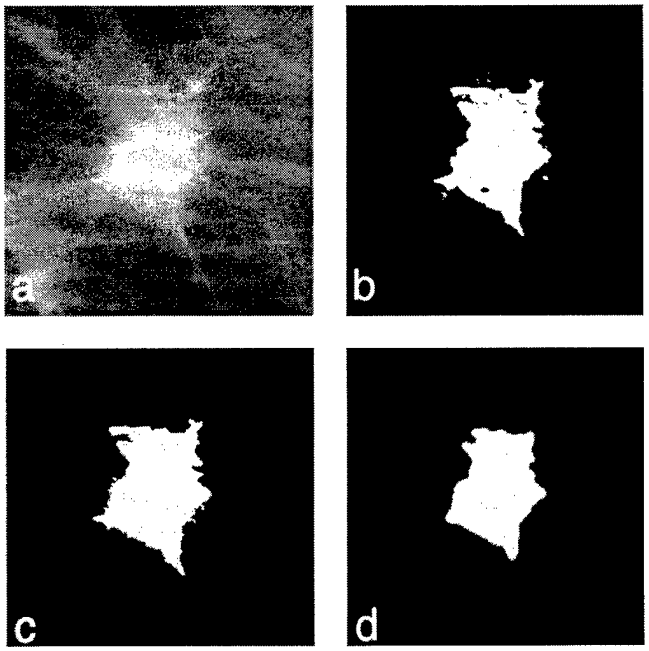


FIG. 2. (a) The original ROI; (b) the result of the initial segmentation; (c) segmented and grown mass object; and (d) smoothed mass object.

object and background tissue. Our segmentation algorithm is described in detail elsewhere.[25,26] Briefly, we obtained several filtered images from the original ROI pixel values, and used the original and filtered pixel values as the elements of a feature vector in the clustering algorithm. The inclusion of spatially filtered images incorporated neighborhood information in the classification of a given pixel.

Figure 2(a) shows an ROI with a spiculated mass. The segmented objects which resulted from the clustering algorithm are shown in Fig. 2(b). After clustering, the largest connected object among all detected objects was selected, filled, and grown in a small region outside its boundary. Details of the region growing algorithm can be found in our previous publications.[25,26] Figure 2(c) shows the result of object selection, filling, and object growing applied to Fig. 2(b). Finally, the borders of the grown object were smoothed by using a morphological opening operation.[27] The opening operation for a binary image consists of the successive application of erosion and dilation operations. In this study, 11×11 and 7×7 pixel circular masks were used for the erosion and dilation operations, respectively. The final smoothed mass object for the ROI in Fig. 2(a) is shown in Fig. 2(d).

In this study, we chose the parameters in the clustering and region growing algorithms such that the mass object was segmented to be slightly smaller than that which could be visually determined on the mammogram. Thus a thin border region along the mass margin was included in the RBST image. Important texture and gradient information at the mass margin was therefore included in the analysis of the region surrounding the mass.

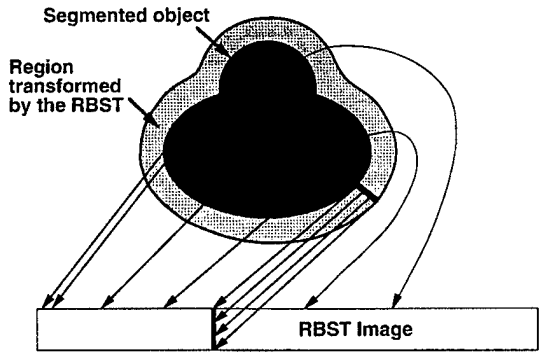


FIG. 3. An illustration of the RBST. The pixels along the object boundary are mapped to the first row of the RBST image. Pixels along a normal to the object boundary are mapped to a column of the RBST image.

C. The RBST image

The RBST maps the pixel values in a band of pixels surrounding the mass onto the Cartesian plane. The mapping designed in this study had the following properties: (i) traversing a closed path at a constant distance from the detected object border in the original image approximately corresponded to moving along a row of the RBST image; and (ii) traveling in a direction normal to the boundary of the detected object in the original image approximately corresponded to moving along a column of the RBST image (Fig. 3). These properties make the RBST well-suited for extracting texture features that radiate from the borders of the mass.

The RBST consists of three main components, edge enumeration, computation of normals, and computation of RBST pixel values. These steps are explained in detail below.

1. Edge enumeration and computation of normals

The border pixels of an object form a closed chain, i.e., starting at an arbitrary pixel, it is possible to move along the chain and return to the starting pixel. Conceptually, the edge enumeration algorithm removes pixels, one at a time, from the edge contour of the object, and places the x and y coordinates of each border pixel on an edge enumeration list. Thus each pixel in the chain is assigned a number, which corresponds to the placement of the pixel in the list.

The algorithm starts by choosing a relatively smooth location on the edge contour, as illustrated in Fig. 4. One pixel (pixel number 1 in Fig. 4) is removed from the edge chain so that the chain is broken. Starting at this break point, pixels are sequentially removed from the chain, and the x and y coordinates of a removed pixel are placed on an enumeration list. Edge enumeration terminates when one returns to the starting pixel after every pixel has been removed form the chain. Since pixel removal is sequential, consecutive pixels in the enumeration list have to be 8-connected neighbors[28] on the edge contour of the object. The algorithm tries to keep the chain in one piece as long as it is possible. Thus referring to Fig. 4, pixel number 12 is followed in the list by pixel number 13, and not pixel number 24. However, when the object shape is complicated, for example, if the object consists of two subobjects joined together with a single bridge

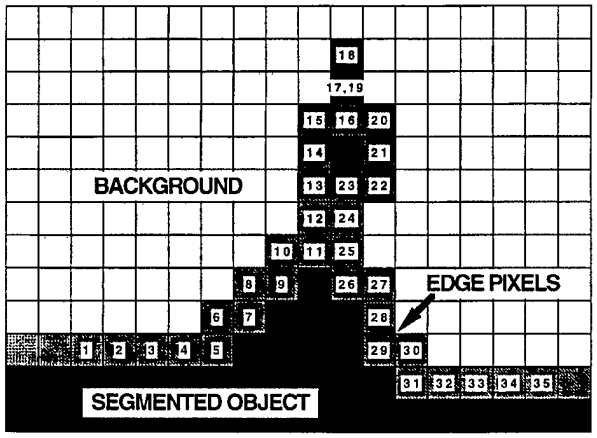


FIG. 4. The edge enumeration algorithm.

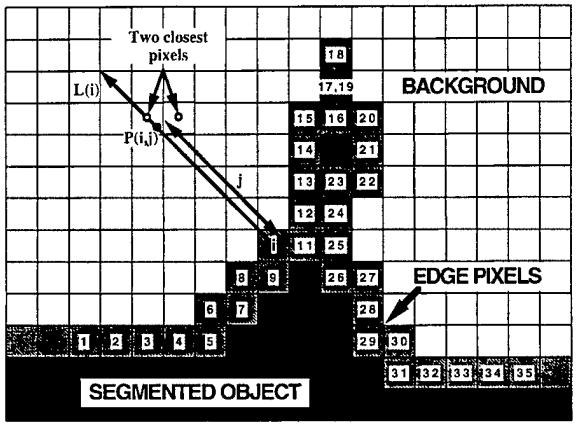


FIG. 6. Computation of RBST pixel values. $p(i,j)$ has a distance j from the pixel i along the normal line $L(i)$. The (i,j)th pixel value in the RBST image is a distance-weighted average of the two closest pixels to $p(i,j)$.

pixel, it may not be possible to keep the chain in one piece. When the chain has to be broken into two, some of the pixels in the chain have to be repeated in the list so that one can return to the starting pixel after removing all the pixels in the chain. The algorithm will then choose a path such that only a small number of pixels in the list are repeated. Thus referring to Fig. 4, pixel number 17 is repeated as pixel number 19 in the list. The number of pixels in the edge enumeration list is denoted as N_e. Since some of the pixels may be repeated in the list, N_e is larger than or equal to the number of edge pixels in the object.

The computation of the normal direction to the object is based on the object shape and the result of the edge enumeration. For a given pixel i in the enumeration list, pixels $i+K$ and $i-K$, occurring K places before and after pixel i are located in the list. The normal direction to the object at edge pixel i is determined as the normal to the line joining edge pixels $i+K$ and $i-K$. This procedure is illustrated graphically in Fig. 5. If $K=1$ as in Fig. 5, only a small neighborhood of a pixel is considered for normal computation, and the computed normals may be noisy. In addition,

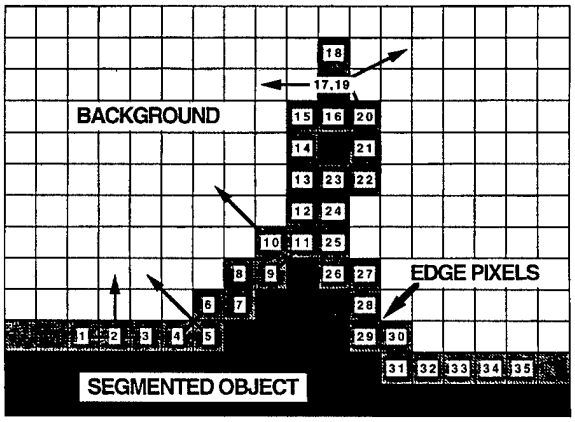


FIG. 5. Computation of normals. For each pixel i, the normal direction $L(i)$ is perpendicular to the line joining pixels $i+K$, $i-K$. For the purpose of illustration, K is set to 1 in this figure. $K=12$ was used in the actual calculation.

$K=1$ confines the number of normal directions to only a small number, since the line joining two neighboring pixels of a given edge pixel can occur in only a small number of directions. On the other extreme, a large value of K may introduce too much smoothing, and some of the fine direction changes in the mass contour may be missed. In this study, it was found experimentally that $K=12$ resulted in a satisfactory normal estimation for most of the mass shapes, and this value was used in the computation of all the RBST images.

2. Computation of RBST pixel values

The basic idea behind the computation of RBST pixel values is as follows. Let $L(i)$ denote the normal to the object at edge pixel i, and let $p(i,j)$ denote the point on the line $L(i)$ which has a distance j from edge pixel i (see Fig. 6.). The value of the pixel in row j, column i of the RBST image is defined as the distance-weighted average of the two closest pixels to $p(i,j)$ in the original image. With this definition, the number of pixels in the enumeration list N_e is equal to the number of columns in the RBST image. The width of the region desired to be transformed determines the number of rows in the RBST image. This definition of the RBST will be referred to as the short RBST.

One difficulty with the short RBST is that as the distance j in the normal direction increases, the length of the closed path surrounding the object at a constant distance j from the object boundary also increases. This may result in undersampling and possibly a loss of information in the RBST image. For example, each of the object border pixels in the original ROI are mapped to the first row of the RBST image. Thus at the first row, transformation from the original image to the RBST image does not result in any information loss. However, when j is large, some pixels in the original image do not contribute to any of the pixels in the RBST image, and the information carried by these pixels will be lost. To reduce the information loss, we increased the number of columns of the RBST image from N_e (defined in the previous paragraph) to $2N_e$. Normals were drawn from each edge

pixel of the object, as well as the midpoints between every two pixels, and the computation of the RBST image was performed as described in the previous paragraph. This definition of the RBST will be referred to as the regular RBST. In this study, we implemented the regular RBST as our main transform. The classification results using the regular RBST are presented in Sec. III.

Depending on the size and shape of the mass, the regular RBST image may contain more pixels than the band of pixels surrounding the mass in the rows adjacent to the segmented mass border. The RBST pixels are computed from the original pixel values using distance-weighted as described above. Therefore, these extra pixels can be considered as the result of an interpolation process. To test whether these extra pixels resulting from interpolation contribute to the performance of the RBST, two options are available. The first option is to interpolate the 256×256 pixel ROI to a larger size by cubic spline interpolation, and to compare the classification accuracy of the texture features extracted from band of pixels in the interpolated image to that of features extracted from the regular RBST image. The second option is to implement the short RBST. The short RBST contains half as many pixels as the regular RBST, and always has fewer pixels than the band of pixels surrounding the mass for convex mass shapes. The classification accuracy using the short RBST can then be compared to that using the original ROI. In this work, we have implemented this second option for comparison, which will simplify RBST implementation if it is found to be as effective as the regular RBST. The results of the comparison are presented in Sec. IV.

Other implementation issues are as follows. A 40-pixel-wide region surrounding the mass object, which corresponds to a 4-mm-wide band, was used to determine the RBST image. The size of the regular RBST image was thus $2N_e$ columns by 40 rows. As discussed in the previous subsection, the distance K used in the computation of normals was 12. For some large masses, some pixels in a 40-pixel-wide band around the mass might fall outside the boundary of the 256×256 pixel ROI. In this study, if $p(i,j)$ fell outside the ROI, the (i,j)th pixel value of the RBST image was flagged as an "invalid" pixel. This in effect reduced the size of the region for extraction of the texture features, as described below. However, since the RBST image of a large mass had a large value of N_e, the reduction in region size did not have a strong effect on the statistical properties of the texture features. An example of an original ROI, segmented mass object, and the RBST image is shown in Fig. 7.

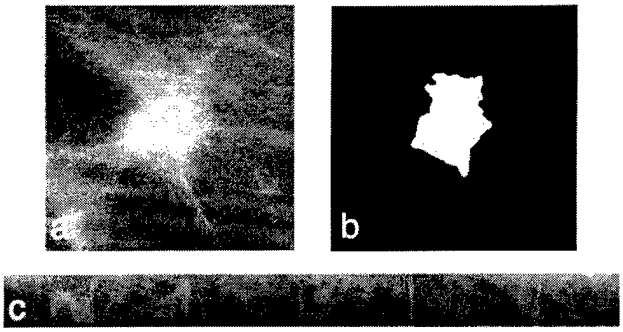


FIG. 7. (a) Original image; (b) segmented mass object; and (c) RBST image.

D. Texture features

The texture features used in this study were calculated from spatial gray-level dependence (SGLD) matrices,[13,14,29] and run-length statistics (RLS) matrices.[30] For comparison purposes, these matrices were computed for three image representations: (i) the entire 256×256 ROI, denoted as $R1$; (ii) a 40-pixel-wide band surrounding the extracted mass boundary, denoted as $R2$; and (iii) the RBST image obtained by applying the RBST to the 40-pixel-wide band, denoted as $R3$. SGLD matrices were constructed from the gray-level images for $R1$, $R2$, and $R3$, and RLS matrices were constructed from the vertical and horizontal gradient images derived from the three image representations, as described below.

1. SGLD texture features

An SGLD matrix can be considered to be a two-dimensional histogram. The element in row r, column c of an SGLD matrix is the joint probability that gray levels r and c occur in a direction θ and at a pixel pair distance of d in the image. The distribution of the SGLD matrix elements reflects the average spatial relationship of pairs of gray-level tones with respect to the distance d and direction θ used in SGLD matrix construction. For example, if the image texture is coarse, and the distance d is small in comparison to the texture element, then pairs with similar gray levels are expected to occur relatively frequently, and pairs with dissimilar gray levels are expected to occur relatively infrequently. Thus the SGLD matrix will be mainly concentrated along the main diagonal. If, in addition, the image is relatively bright (indicated by high pixel values), then the SGLD matrix will be concentrated around the lower main diagonal. SGLD texture features, described in the next paragraph, extract this information from the SGLD matrix.

Based on our previous studies,[13] a bit depth of eight bits was used in the SGLD matrix construction, i.e., the least significant four bits of the 12-bit pixel values were discarded. Eight texture measures, namely, correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, and inertia were extracted from each SGLD matrix at eight different pixel pair distances, ($d=1$, 2, 3, 4, 6, 8, 12, and 16) and in four directions ($\theta=0°$, 45°, 90°, and 135°). Therefore, a total of 256 SGLD features were calculated for each image representation. The formulation of these texture measures has been described in the literature.[13,14,29] These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.[29] For example, the energy feature, which is the sum-of-squares of the SGLD matrix elements, is smallest when all the elements of the SGLD matrix are equal, i.e., when all the pixel pairs occur with equal probability. This would indicate that the image does not have a

lot of structure. As another example, the inertia feature, which is the moment of inertia of the SGLD matrix around its main diagonal, measures the spread of the matrix elements around the main diagonal. A high value of this feature means that the spread is high, which indicates that the size of the image texture elements are comparable to, or smaller than the pixel pair distance d. Although such examples provide an idea about the meaning of these features, it is difficult to establish a one-to-one correspondence between the qualitative image characteristics and the extracted features.[29]

In this study, special care was taken in the construction of the SGLD matrix, since some of the image representations contained invalid pixel values as specified in the previous subsection. When the SGLD matrices were constructed, pixel pairs involving invalid pixel values were not accumulated in the SGLD matrix.

2. RLS texture features

RLS texture features were extracted from vertical and horizontal gradient magnitude images, which were obtained by filtering the image representation of interest by the horizontally or vertically oriented Sobel filters, and computing the absolute value of the filtered image.

A gray-level run is a set of consecutive, colinear pixels in a given direction which have the same gray-level value. A run length is the number of pixels in a run.[30] The RLS matrix describes the run-length statistics for each gray-level value in the image. The element in row r, column c of an RLS matrix is the number of times that the gray level r in the image possesses a run length of c in a given direction.

Analogous to SGLD matrix computation, invalid pixel values were excluded from the RLS matrix computation. If a large bit depth is used in RLS matrix computation, the resulting run lengths are very short for all of the images, and the discriminatory power may not be high. Conversely, if the bit depth is too small, then run lengths become predominantly long. In this study, it was found experimentally that a bit depth of 5 bits in RLS matrix computation resulted in a good compromise.

Five texture measures, namely, short runs emphasis, long runs emphasis, gray-level nonuniformity, run-length nonuniformity, and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta=0°$, and $\theta=90°$. Therefore, a total of 20 RLS features were calculated for each image representation. The definition of the RLS texture measures used in this study can be found in the literature.[30] It is possible to crudely describe the dependence of these features on the image characteristics, e.g., the run percentage feature value is small for images with long linear structures, and the gray-level nonuniformity feature value is small for images where runs are equally distributed throughout the gray levels. However, it is again difficult to establish a one-to-one correspondence between the qualitative image characteristics and the extracted features.

E. Classification

Linear discriminant analysis[31,32] was used to classify malignant and benign masses based on the extracted texture features. A stepwise feature selection procedure with the minimization of Wilks' lambda (the ratio of within-group sum of squares to the total sum of squares) was used as the optimization criterion to select effective predictor variables. Stepwise feature selection is an iterative procedure, where one feature is entered into or removed from the selected feature pool at each step by analyzing its effect on the selection criterion. In the feature entry phase of a step, the available features are entered into the selected feature pool one at a time. The significance of the change in Wilks' lambda, as measured by F-statistics, when a feature is entered into the selected feature pool is compared to a threshold F_{in}. The feature with the highest significance is entered into the selected feature pool only if the significance is higher than F_{in}. Likewise, in the feature removal phase, features that were already selected are removed from the selected feature pool one at a time, and the significance of change in Wilks' lambda is compared to a threshold F_{out}. The feature with the least significance is removed from the selected feature pool only if the significance is lower than F_{out}. Since the optimal values of the F_{in} and F_{out} parameters are not known *a priori*, we varied both parameters, and tried to obtain the feature combinations that yielded the highest classification accuracy for each of the three image representations. Details about the application of stepwise linear discriminant analysis to CAD can be found in our previous publications.[13,14,26]

A leave-one-case-out method was used to train and test the classifier. In this method, all films belonging to one patient were left out from the classifier design group at the same time. A linear discriminant function was formed using the design group, and test discriminant scores were computed for the left-out films using the linear discriminant function. This process cycled through the data set until every patient's films were used as test films once. The test discriminant scores of all films were analyzed using receiver operating characteristic (ROC) methodology[33] to evaluate the classifier performance. The discriminant scores of the malignant and benign masses were used as the decision variable in the LABROC1 program,[34] which provided the ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area A_z under the ROC curve. The CLABROC program[35] was used to test the statistical significance of the difference between pairs of ROC curves obtained using texture features extracted from $R1$, $R2$, and $R3$ under corresponding conditions.

F. Computational considerations

Segmentation, image transformation, feature extraction, and classifier design steps of our algorithm were executed on an AlphaStation 500 (400-MHz Alpha chip), and the feature selection step was performed on a PC compatible computer with a 90-MHz Pentium processor. The classification for the entire data set of 168 images took less than an hour, which meant that the classifier design and classification for each

TABLE I. Classifier performance with SGLD texture features, extracted from (a) $R1$ (the original ROI), (b) $R2$ (the 40-pixel-wide region surrounding the mass), and (c) $R3$ (the RBST image). F_{in} and F_{out} values are thresholds used in the stepwise feature selection method for entering and removing features from the selected feature pool. In general, lower thresholds result in a larger number of selected features.

(a)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
1.2	1.0	13	0.85±0.03	0.77±0.04
1.1	1.3	17	0.87±0.03	0.79±0.03
1.1	1.2	20	0.88±0.03	0.78±0.04
0.8	0.6	25	0.91±0.02	0.81±0.03*
0.6	0.4	26	0.91±0.02	0.79±0.03

(b)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
0.6	0.8	5	0.78±0.03	0.74±0.04
0.73	0.73	17	0.85±0.03	0.79±0.03
0.7	0.7	21	0.90±0.02	0.83±0.03
0.6	0.4	32	0.96±0.01	0.87±0.03*
0.4	0.2	34	0.96±0.01	0.86±0.03

(c)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
2.4	2.2	9	0.92±0.02	0.89±0.03
2.2	2.0	12	0.94±0.02	0.91±0.02*
0.6	0.4	18	0.95±0.02	0.90±0.02
0.4	0.2	23	0.95±0.02	0.89±0.02

TABLE II. Classifier performance with RLS texture features, extracted from (a) $R1$ (the original ROI), (b) $R2$ (the 40-pixel-wide region surrounding the mass), and (c) $R3$ (the RBST image). F_{in} and F_{out} values are thresholds used in the stepwise feature selection method for entering and removing features from the selected feature pool. In general, lower thresholds result in a larger number of selected features.

(a)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
1.6	1.4	3	0.74±0.04	0.70±0.04
1.4	1.2	4	0.74±0.04	0.70±0.04
1.2	1.0	5	0.75±0.03	0.70±0.04*
0.2	0.1	9	0.75±0.03	0.67±0.04

(b)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
1.2	1.0	2	0.73*m*0.04	0.71±0.04
0.8	0.6	5	0.76±0.04	0.72±0.04*
0.6	0.4	6	0.77±0.04	0.72±0.04

(c)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
5.2	5.0	5	0.86±0.03	0.84±0.03*
3.8	2.7	6	0.87±0.03	0.83±0.03
1.2	1.0	7	0.87±0.03	0.84±0.03
1.0	0.8	9	0.88±0.03	0.83±0.03

mass was performed in less than 30 s. If a trained classifier is implemented, the feature selection and classifier design steps will not be needed for classifying an unknown case, and the computation time will be shorter.

III. RESULTS

In this section, we present classification results with texture features derived from the $R1$, $R2$, and $R3$ image representations. Since the optimal number of features is not known *a priori*, we varied the F_{in} and F_{out} parameters in the stepwise linear discriminant analysis and tried to obtain a range in the number of selected features for each image representation. The F_{in} and F_{out} values, as well as the number of features are tabulated for different conditions in the following subsections. After feature selection and classifier design were completed, each designed classifier was applied to its design samples, and a training A_z value was obtained. Since our database contained images from 72 different patients, 72 classifiers were trained for each feature combination in a leave-one-case-out paradigm. The training A_z values and their standard deviations in the following tables represent the averages of these quantities from the output of the LABROC1 program over the 72 classifiers. After training and testing were completed on all of the films for a feature combination, the test A_z and its standard deviation were estimated by the LABROC1 program using the test scores.

A. SGLD feature space

Tables I(a)–I(c) show the training and test classification accuracies using the SGLD features derived from the $R1$, $R2$, and $R3$ image representations, respectively. The highest test classification result for each representation is marked with an asterisk. It can be observed that the range of selected features for each representation was large enough so that the maximum occurred within the range, and not at the highest or lowest number of selected features. The test classification results in Table I(a), as well as those in Table I(c) were within one standard deviation of each other. The results in Table I(b) had a larger variation, due to the wider range in the number of selected features. The difference between the best classification results using $R1$ and $R3$ was statistically significant ($p<0.03$). The difference between the best classification results using $R2$ and $R3$ did not achieve statistical significance. The texture features that were selected most frequently in the SGLD feature space were difference entropy and inverse difference moment. Both of these features measure the spread of the SGLD matrix along lines parallel to the main diagonal. Therefore, they are measures of the local nonhomogeneity of the image.

B. RLS feature space

Tables II(a)–II(c) show the training and test classification accuracies using the RLS features derived from the $R1$, $R2$, and $R3$ image representations, respectively. The highest test classification results are marked with an asterisk. Since we had only 20 RLS texture features, the variation in the number of features in each table was smaller compared to that for the SGLD texture features. The test classification results within each table were again within one standard deviation of each other. The difference between the best classification results using $R1$ and $R3$, as well as $R2$ and $R3$ were statistically

TABLE III. Classifier performance with combined texture features, extracted from (a) $R1$ (the original ROI), (b) $R2$ (the 40-pixel-wide region surrounding the mass), and (c) $R3$ (the RBST image). F_{in} and F_{out} values are thresholds used in the stepwise feature selection method for entering and removing features from the selected feature pool. In general, lower thresholds result in a larger number of selected features.

(a)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
1.8	1.6	8	0.82±0.03	0.77±0.04
1.35	1.2	14	0.87±0.03	0.80±0.03
1.3	1.2	16	0.88±0.03	0.81±0.03*
1.2	1.0	19	0.90±0.02	0.79±0.03
1.0	0.8	20	0.90±0.02	0.78±0.03
0.8	0.6	30	0.92±0.02	0.80±0.03

(b)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
1.8	1.6	15	0.91±0.02	0.86±0.03
1.4	1.2	19	0.93±0.02	0.86±0.03
1.2	1.0	20	0.93±0.02	0.87±0.03*
1.1	1.1	21	0.93±0.02	0.86±0.03
1.0	0.8	25	0.94±0.02	0.86±0.03
0.8	0.8	27	0.94±0.02	0.85±0.03

(c)

F_{in}	F_{out}	Num. of features	Training A_z	Test A_z
3.0	2.8	11	0.92±0.02	0.89±0.02
2.6	2.4	14	0.96±0.01	0.94±0.02
2.2	2.0	18	0.97±0.01	0.94±0.02
1.6	1.4	20	0.98±0.01	0.94±0.02*
1.0	1.0	22	0.97±0.01	0.93±0.02

significant ($p<0.01$ for both). Long runs emphasis and short runs emphasis were the two features that were selected most frequently in the RLS feature space. These features emphasize long and short runs in the image, and therefore indicate the existence of long or short linear structures in the image, respectively.

C. Combined SGLD and RLS feature space

Tables III(a)–III(c) show the training and test classification accuracies using both the SGLD features and the RLS features derived from the $R1$, $R2$, and $R3$ image representations, respectively. In analogy to SGLD feature selection, the range of selected features in this subsection was large enough so that the maximum occurred within the range. Almost all of the test classification results within each table were within one standard deviation of each other. The ROC curves for the classifiers with the highest test accuracy, marked by an asterisk in the tables, are plotted in Fig. 8. The difference between the best classification results using $R1$ and $R3$, as well as $R2$ and $R3$ were again statistically significant ($p<0.01$ for both). The distribution of the test discriminant scores obtained by using features extracted from the RBST images is shown in Fig. 9. By choosing an appropriate decision threshold on the test discriminant scores, more than 30% of the benign masses could correctly be identified without missing any malignant masses. Difference entropy, inverse difference moment, and long runs emphasis were the three features that were selected most frequently in the combined feature space.

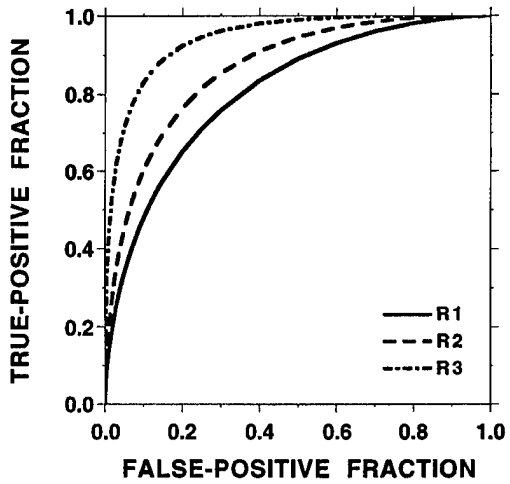


FIG. 8. ROC curves for $R1$ (the original ROI), $R2$ (the 40-pixel-wide region surrounding the mass), and $R3$ (the RBST image). Classification was performed in the combined SGLD and RLS feature space.

IV. DISCUSSION

We have designed and implemented a new rubber band straightening transform, and used this transformation for classifying malignant and benign breast masses. Our results showed that both SGLD features and RLS features, as well as the combined feature set extracted from the RBST images ($R3$) were significantly more effective than similar features extracted from the entire 256×256 ROI containing the mass ($R1$). The RBST image was obtained by transforming a 40-pixel (4 mm) wide band surrounding the segmented mass. For this reason, we compared the classification effectiveness of texture features extracted from a 40-pixel-wide band surrounding the segmented mass ($R2$) with those from the RBST image ($R3$). Our results showed that RLS features extracted from $R3$ were significantly more effective than RLS features extracted from $R2$. The classification accuracy using SGLD features extracted from $R3$ was also higher

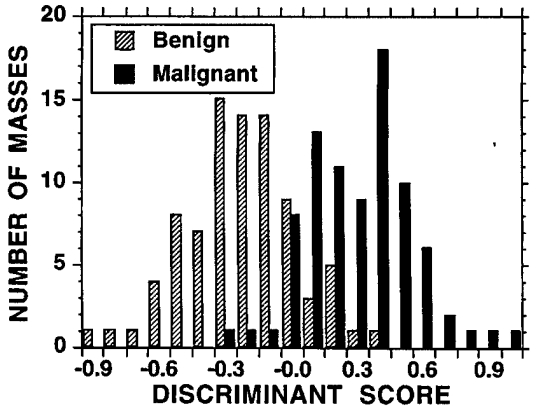


FIG. 9. The distribution of the test discriminant scores obtained by using combined SGLD and RLS features extracted from $R3$ (the RBST images).

than $R2$, although the difference did not achieve statistical significance. In the combined feature space, we again observed significantly higher classification accuracy with the use of the RBST images.

It is expected that the texture of the region surrounding a mass has a radial dependence, because possible speculations and the gradient of the opacity caused by the mass are approximately radially oriented. However, most texture extraction methods are designed for texture orientations in a uniform direction (horizontal, vertical, or at a certain angle between these two directions). By transforming the region surrounding a mass into an RBST image, we have attempted to create a transformed image in which texture orientations become more suitable for feature extraction using existing techniques. The results of this study indicate that our approach is promising.

The width of the region transformed by the RBST was selected as 40 pixels (4 mm) in this paper. In another publication on classification of masses,[36] the same size was used inside and outside the mass for feature extraction. If the width of this band is too small, then the RBST image may exclude some of the border regions with useful texture in the original image. If the width is too large, then the statistical feature variations of the structures far away from the mass, which carry little or no information on its probability of malignancy, may be included and degrade the classifier performance. We did not perform a systematic study of the effect of the size of this region on the classification accuracy. However, to test whether this size was a critical parameter, we obtained RBST images for 30- and 50-pixel-wide bands, and extracted the same set of features as discussed in Sec. II from these images. With 30- and 50-pixel-wide bands, the test classification accuracy A_z using the combined feature space was 0.93 and 0.92, respectively. The difference between these results and the best result in Table III(c) ($A_z=0.94$) was not statistically significant. We therefore surmise that the classification accuracy will not be very sensitive to this size. It is reasonable to expect that the size of the region surrounding the mass that contains useful information about its malignancy will change with the size of the mass. Therefore, one may improve the classification results obtained in this paper by adaptively changing the size of the region transformed by the RBST depending on the size of the mass. This will be investigated in the future.

The length of the RBST image in this paper was $2N_e$ pixels, where N_e is the number of edge pixels of the segmented mass. Depending on the size and shape of the mass, the RBST image thus defined may contain more pixels than the 40-pixel-wide band area surrounding the mass. To test whether these extra pixels contribute to the performance of the RBST, we implemented a variation of the RBST termed the short RBST, which produces an RBST image having a length of N_e pixels. For a convex mass shape, the short RBST image will always have fewer pixels than the band of pixels surrounding the mass.

After the computation of the short RBST images, feature extraction, selection, and classification were performed in the same way as the regular images, as discussed in Secs. II and III. The test A_z scores using the SGLD, RLS, and combined feature spaces were 0.91, 0.81, and 0.93, respectively. These results are equal to, or slightly worse than the best test results in Tables I(c), II(c), and III(c) marked with an asterisk. The difference between the A_z values obtained using the corresponding feature spaces was not statistically significant. The statistical differences between the classification results obtained using the short RBST and the $R1$ or $R2$ image representations were similar to the differences between the regular RBST and the $R1$ or $R2$ image representations. More precisely, the classification results obtained using the short RBST were significantly better than those obtained using both $R1$ and $R2$ representations in the RLS and combined feature spaces ($p<0.05$). In the SGLD feature space, the difference between the classification results using the short RBST and the $R1$ image representation was statistically significant ($p<0.05$), but the difference between the short RBST and the $R2$ image representation did not achieve statistical significance. These results show that the extra pixels resulting from the interpolation in the computation of the regular RBST do not provide an advantage to the RBST over the other image representations. This is consistent with the expectation that interpolation generally does not increase image information.

The test A_z values obtained from a given representation in a given feature space were within one standard deviation of each other. This meant that the optimal values of F_{in} and F_{out}, and therefore the number of selected features, were not critical for designing the classifiers. However, the feature selection process itself is a critical component in classification, as shown in our previous study.[26] In many of the tables, one can observe the so-called peaking phenomenon,[37] which means that when a moderate number of design samples is available for classifier design, the test accuracy first increases, but later starts to decrease as the number of features is increased.

As discussed in Sec. II, the probability of malignancy of each mass, based on the mammographic appearance, was ranked by a radiologist experienced in mammography (Fig. 1). Based on this ranking, an ROC curve was estimated using the LABROC1 program, and plotted in Fig. 10. The figure also plots the ROC curve obtained by using the combined texture features extracted from the RBST images. The A_z value obtained by the malignancy ranking of the radiologist was 0.89 ± 0.03. The difference between the ROC curves using the computerized classification algorithm ($A_z=0.94\pm0.02$) and the malignancy rating of the radiologist was statistically significant (two-tailed p level$=0.03$). This result also highlights the promise of our approach.

In this study, the ranking by the radiologist, as well as the computer scores, were based only on the appearance of the mass on a single mammogram. Other views of the patient, such as different views of the same breast, films of the other breast, previous mammograms, spot, and magnification views were not used to assist either the radiologist or the computer. Therefore, the discussion in the previous paragraph only compares the performances of the radiologist and the computer under specific laboratory conditions. The ma-

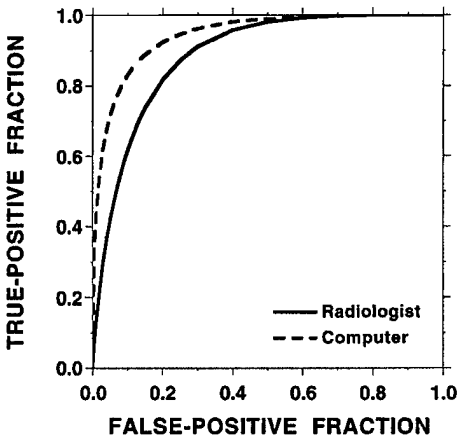


FIG. 10. ROC curves obtained by using the radiologist's malignancy rating ($A_z = 0.89 \pm 0.03$) and the computer's discriminant score output ($A_z = 0.94 \pm 0.02$) with features extracted from $R3$ (the RBST images).

lignant and benign classification by radiologists can be expected to be more accurate when different views of the same mass are examined. The accuracy of computerized characterization is also expected to improve when the features or discriminant scores obtained from different mammograms of the same patient are combined. However, this was not performed in this study since our purpose was to compare the usefulness of the RBST with other image representations. Similarly, the ROC curves and the A_z scores in Sec. III do not necessarily reflect the accuracy expected to be obtained under clinical conditions, but they show the trend that the RBST is useful.

The segmentation, feature extraction and classification methods used in this work and that of Huo *et al.*[23] are different. However, in both investigations, features extracted from the area surrounding the segmented mass resulted in better classification accuracy compared to features extracted from other regions. Since the data sets are different, it is difficult to compare the performances of the two methods. The data set used in our study was almost twice as large as that used by the other study.[23] Huo *et al.* used an *ad hoc* method for geometric shape correction, and employed the maximum of the corrected measure in four different neighborhoods for better classification results. It remains to be seen whether these methods are generalizable to larger data sets. Similarly, when our feature selection and classification methods are applied to a larger data set, the selected features and the coefficients of the selected features in linear discriminant analysis are likely to change. It remains to be seen whether the classification accuracy will decrease under these conditions.

An advantage of our approach compared to some recent publications[22,23] is that the mass characterization method proposed in this study is applicable to both spiculated and nonspiculated masses. As summarized in Table IV, at a 95% overall sensitivity level, our algorithm was able to correctly diagnose 100% of the spiculated malignant masses, and 89% of the nonspiculated malignant masses. At the same overall sensitivity level, the radiologist's rankings also showed 100% and 89% true-positive rates for spiculated malignant and nonspiculated malignant masses, respectively. However, at this sensitivity level, the computer had a 81% specificity (69 true negatives—68 nonspiculated and 1 spiculated) and the radiologist had a 60% specificity (51 true negatives—50 nonspiculated and 1 spiculated).

V. CONCLUSION

We have developed a new image transformation method, referred to as RBST, for the characterization of mammographic masses. The results of our classification study indicate that texture features extracted from the transformed images are useful in differentiation of malignant and benign masses. With the best combination of texture features, the test A_z value on our database of 168 mammograms reached 0.94. It was found that texture features extracted from the transformed images were significantly more effective than features extracted from the ROIs before the transformation. This demonstrates the usefulness of the RBST. Before the applicability of our approach can be tested in a clinical setting, further studies need to be performed with a larger database to investigate the generalizability of these results. The combination of information from mammograms of different views obtained from the same patient will be investigated. The combination of texture and morphological features for benign and malignant characterization of masses will also be studied.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award (B.S.) from the USAMRMC (DAMD 17-96-1-6012) and a USPHS Grant CA 48129. The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC1 and CLABROC programs.

TABLE IV. Sensitivity (Sens.) and specificity (Spec.) for spiculated (S), and nonspiculated (NS) masses at a 95% overall sensitivity level.

	Malignant ($n=83$)			Benign ($n=85$)		
	S ($n=45$)	NS ($n=38$)	Overall Sens. ($n=83$)	S ($n=6$)	NS ($n=79$)	Overall Spec. ($n=85$)
Computer	100% ($n=45$)	89% ($n=34$)	95% ($n=79$)	17% ($n=1$)	86% ($n=68$)	81% ($n=69$)
Radiologist	100% ($n=45$)	89% ($n=34$)	95% ($n=79$)	17% ($n=1$)	63% ($n=50$)	60% ($n=51$)

[1]H. C. Zuckerman, "The role of mammography in the diagnosis of breast cancer," in *Breast Cancer, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw-Hill, New York, 1987), pp. 152–172.
[2]D. B. Kopans, "The positive predictive value of mammography," Am. J. Roentgenol. **158**, 521–526 (1992).
[3]D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," Curr. Opin. Radiol. **4**, 123–129 (1992).
[4]M. Moskowitz, "Impact of a priori medical decisions on screening for breast cancer," Radiology **171**, 605–608 (1989).
[5]C. J. D'Orsi, D. J. Getty, J. A. Swets, R. M. Pickett, S. S. Seltzer, and B. J. McNeil, "Reading and decision aids for improved accuracy and standardization of mammographic diagnosis," Radiology **184**, 619–622 (1992).
[6]G. Hermann, C. Janus, I. S. Schwartz, B. Krivisky, S. Bier, and J. G. Rabinowitz, "Nonpalpable breast lesions: Accuracy of prebiopsy mammographic diagnosis," Radiology **165**, 323–326 (1987).
[7]F. M. Hall, J. M. Storella, D. Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography," Radiology **167**, 353–358 (1988).
[8]H. G. Jacobson and J. Edeiken, "Biopsy of occult breast lesions: Analysis of 1261 abnormalities," JAMA **263**, 2341–2343 (1990).
[9]F. F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," Invest. Radiol. **28**, 473–481 (1993).
[10]W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).
[11]H. D. Li, M. Kallergi, L. P. Clarke, V. K. Jain, and R. A. Clark, "Markov random field for tumor detection in digital mammography," IEEE Trans. Med. Imaging **14**, 565–576 (1995).
[12]B. Zheng, Y.-H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis," Acad. Radiol. **2**, 959–966 (1995).
[13]H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).
[14]D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," Med. Phys. **22**, 1501–1513 (1995).
[15]N. Petrick, H.-P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," IEEE Trans. Med. Imaging **15**, 59–67 (1996).
[16]J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon," Radiology **196**, 817–822 (1995).
[17]D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," Invest. Radiol. **23**, 240–252 (1988).
[18]Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," Radiology **187**, 81–87 (1993).
[19]D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammography," IEEE Trans. Med. Imaging **9**, 233–241 (1990).
[20]J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).
[21]M. L. Giger, C. J. Vyborny, and R. A. Schmidt, "Computerized characterization of mammographic masses: Analysis of spiculation," Cancer Lett. **77**, 201–211 (1994).
[22]S. Pohlman, K. A. Powell, N. A. Obuchowski, W. A. Chilcote, and S. G. Broniatowski, "Quantitative classification of breast tumors in digitized mammograms," Med. Phys. **23**, 1337–1345 (1996).
[23]Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," Med. Phys. **22**, 1569–1579 (1995).
[24]L. Tabar and P. B. Dean, *Teaching Atlas of Mammography* (Thieme, New York, 1985).
[25]B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: An artificial neural network with morphological features," in *Proceedings of the World Congress on Neural Networks* (INNS Press, New Jersey, 1995), pp. 876–879.
[26]B. Sahiner, H.-P. Chan, D. Wei, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," Med. Phys. **23**, 1671–1684 (1996).
[27]J. Serra, *Image Analysis and Mathematical Morphology* (Academic, London, 1982).
[28]R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
[29]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Trans. Syst. Man Cybern. **3**, 610–621 (1973).
[30]M. M. Galloway, "Texture analysis using gray level run lengths," Comput. Graph. Image Process. **4**, 172–179 (1975).
[31]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).
[32]M. J. Norusis, *SPSS Professional Statistics 6.1* (SPSS, Chicago, 1993).
[33]C. E. Metz, "ROC methodology in radiographic imaging," Invest. Radiol. **21**, 720–733 (1986).
[34]C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously distributed test results," presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim, CA (1990).
[35]C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance of differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging: Proceedings of the 8th Conference*, edited by F. Deconinck (Martinus Nijhoff, Boston, Brussels, 1984), pp. 432–445.
[36]R. M. Rangayyan, N. El-Faramawy, J. E. L. Desautels, and O. A. Alim, "Discrimination between benign and malignant breast tumors using a region-based measure of edge profile acutance," in *Digital Mammography '96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996), pp. 213–218.
[37]R. F. Wagner, D. G. Brown, J-P. Guedon, K. J. Myers, and K. A. Wear, "Multivariate Gaussian pattern classification: Effects of finite sample size and the addition of correlated or noisy features on summary measures of goodness," in *Information Processing in Medical Imaging*, edited by H. H. Barrett and A. F. Gmitro (Springer-Verlag, Berlin, 1993), pp. 507–524.

Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces

Heang-Ping Chan[a] and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

Kwok Leung Lam
Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan 48109

Nicholas Petrick, Mark A. Helvie, Mitchell M. Goodsitt, and Dorit D. Adler
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 24 September 1997; accepted for publication 20 July 1998)

We are developing computerized feature extraction and classification methods to analyze malignant and benign microcalcifications on digitized mammograms. Morphological features that described the size, contrast, and shape of microcalcifications and their variations within a cluster were designed to characterize microcalcifications segmented from the mammographic background. Texture features were derived from the spatial gray-level dependence (SGLD) matrices constructed at multiple distances and directions from tissue regions containing microcalcifications. A genetic algorithm (GA) based feature selection technique was used to select the best feature subset from the multi-dimensional feature spaces. The GA-based method was compared to the commonly used feature selection method based on the stepwise linear discriminant analysis (LDA) procedure. Linear discriminant classifiers using the selected features as input predictor variables were formulated for the classification task. The discriminant scores output from the classifiers were analyzed by receiver operating characteristic (ROC) methodology and the classification accuracy was quantified by the area, A_z, under the ROC curve. We analyzed a data set of 145 mammographic microcalcification clusters in this study. It was found that the feature subsets selected by the GA-based method are comparable to or slightly better than those selected by the stepwise LDA method. The texture features ($A_z=0.84$) were more effective than morphological features ($A_z=0.79$) in distinguishing malignant and benign microcalcifications. The highest classification accuracy ($A_z=0.89$) was obtained in the combined texture and morphological feature space. The improvement was statistically significant in comparison to classification in either the morphological ($p=0.002$) or the texture ($p=0.04$) feature space alone. The classifier using the best feature subset from the combined feature space and an appropriate decision threshold could correctly identify 35% of the benign clusters without missing a malignant cluster. When the average discriminant score from all views of the same cluster was used for classification, the A_z value increased to 0.93 and the classifier could identify 50% of the benign clusters at 100% sensitivity for malignancy. Alternatively, if the minimum discriminant score from all views of the same cluster was used, the A_z value would be 0.90 and a specificity of 32% would be obtained at 100% sensitivity. The results of this study indicate the potential of using combined morphological and texture features for computer-aided classification of microcalcifications. [S0094-2405(98)00910-9]

Key words: computer-aided diagnosis, mammography, microcalcifications, genetic algorithm, linear discriminant analysis, ROC analysis

I. INTRODUCTION

Mammography is the most sensitive method for early detection of breast cancers. However, its specificity for differentiating malignant and benign lesions is relatively low. In the United States, the positive predictive value of mammography ranges from about 15% to 30%.[1,2] Various methods are being developed to improve the sensitivity and specificity of breast cancer detection.[3] Computer-aided diagnosis (CAD) is considered to be one of the promising approaches that may improve the efficacy of mammography.[4] Properly designed CAD algorithms can automatically detect suspicious lesions on a mammogram and alert the radiologist to these regions. They can also extract image features from regions of interest (ROIs) and estimate the likelihood of malignancy for a given lesion, thereby providing the radiologist with additional information for making diagnostic decisions.

There are two major approaches to the development of CAD schemes for classification of mammographic abnormalities. One approach uses computer vision techniques to extract image features from the digitized mammograms and classify the lesions based on the computer-extracted features. The computer-extracted features can include morphological features that are commonly used by radiologists for diagno-

sis, as well as texture features that may not be readily perceived by human eyes. The computerized analysis may therefore increase the utilization of mammographic image information and improve the accuracy of differentiating malignant and benign lesions. The other approach uses radiologists' ratings of mammographic features or encodes the radiologists' readings with numerical values. The lesions are then classified based on these radiologist-extracted features. This approach assists radiologists by systematically extracting image features and by optimally merging the features with a statistical classifier to reach a diagnostic decision. Additional risk factors based on patient demographic information and medical or family histories may also be included as input in either approach.

A number of investigators have developed feature extraction and classification methods for characterization of mammographic masses or microcalcifications. Ackerman *et al.*[5] developed 4 measures of malignancy and classified lesions recorded on 120 digitized xeroradiographs by 3 decision methods. Kilday *et al.*[6] used 7 shape descriptors and patient age to classify 39 masses and could correctly classify 69% of the masses. Huo *et al.*[7] analyzed the spiculation of masses using a radial edge-gradient analysis technique and achieved an area, A_z, under the receiver operating characteristic (ROC) curve of 0.88 in a data set of 95 masses. Sahiner *et al.*[8,9] developed a rubber-band straightening image transformation technique to analyze the texture in the region surrounding a mass and obtained an A_z of 0.94 in a data set of 168 masses. Pohlman *et al.*[10] extracted 6 morphological descriptors to classify 47 masses and obtained A_z values ranging from 0.76 to 0.93. Wee *et al.*[11] analyzed 51 microcalcification clusters on specimen radiographs using the average gray level, contrast, and horizontal length of the microcalcifications and obtained 84% correct classification. Fox *et al.*[12] included cluster features in their classifier and obtained 67% correct classification in a data set of 100 clusters from specimen radiographs. Chan *et al.*[13–18] developed morphological and texture features and evaluated various feature classifiers for differentiation of malignant and benign microcalcifications. Shen *et al.*[19] used 3 shape features, compactness, moments, and Fourier descriptors to classify 143 individual microcalcifications with a nearest neighbor classifier and obtained 100% classification accuracy. Wu *et al.*[20] classified 80 pathologic specimens radiographs with a convolution neural network and obtained an A_z of 0.90. Jiang *et al.*[21] trained a neural network classifier to analyze 8 features extracted from microcalcification clusters and obtained an A_z of 0.92 in a data set of 53 patients. Thiele *et al.*[22] extracted texture and fractal features from the tissue region surrounding a microcalcification cluster for classification and achieved a sensitivity of 89% at a specificity of 83% for 54 clusters. Dhawan *et al.*[23] used features derived from first-order and second-order gray-level histogram statistics and obtained an A_z of 0.81 with a neural network classifier for a data set of 191 clusters.

Computerized classification of mammographic lesions using radiologist-extracted features has also been reported by a number of investigators. Ackerman *et al.*[24] estimated the probability of malignancy of mammographic lesions by analyzing 36 radiologist-extracted characteristics with an automatic clustering algorithm and obtained a specificity of 45% at a sensitivity of 100% in a data set of 102 cases. Gale *et al.*[25] analyzed 12 radiologist-extracted features of mammographic lesions with a computer algorithm and obtained a specificity of 88% at a sensitivity of 79% in a data base of 500 patients. Getty *et al.*[26] developed a computer classifier to enhance the differentiation of malignant and benign lesions by a radiologist during interpretation of xeromammograms. Using a similar approach, D'Orsi *et al.*[27] evaluated a computer aid and obtained an improvement of about 0.05 in sensitivity or specificity in mammographic reading. Wu *et al.*[28] trained a neural network to merge 14 radiologist-extracted features for classification of mammographic lesions and obtained an A_z of 0.89. Baker *et al.*[29] trained a neural network based on the lexicon of the Breast Imaging Recording and Data System of the American College of Radiology and found that the neural network could improve the positive predictive value from 35% to 61% in 206 lesions. Lo *et al.*[30] used a similar approach to predict breast cancer invasion and obtained an A_z of 0.91 for 96 lesions. Although the results of these studies varied over a wide range and the performances of the computer algorithms are expected to depend strongly on data set, they indicate the potential of using CAD techniques to improve the diagnostic accuracy of differentiating malignant and benign lesions.

In our early studies, we found that texture features extracted from spatial gray-level dependence (SGLD) matrices at multiple distances were useful for differentiating malignant and benign masses on mammograms. This may be attributed to the texture changes in the breast tissue due to a developing malignancy. The usefulness of SGLD texture measures in differentiating malignant and benign breast tissues was further demonstrated by analysis of mammographic microcalcifications.[17,18,31] In a preliminary study, we developed morphological features to describe the size, shape, and contrast of the individual microcalcifications and their variation within a cluster. We used these features to classify the microcalcifications and obtained moderate results.[13,15] In the present study, we expanded the data set and explored the feasibility of combining texture and morphological features for classification of microcalcifications. The classification accuracy in the combined feature space was compared with those obtained in the texture feature space or in the morphological feature space alone. We also studied the use of a genetic algorithm[32–34] (GA) to select a feature subset from the large-dimension feature spaces, and compared the classification results to those obtained from features selected with stepwise linear discriminant analysis (LDA).[35] Linear discriminant classifiers[36] were designed for the classification tasks. The performance of the classifiers was analyzed with ROC methodology[37] and the classification accuracy was quantified with the area, A_z, under the ROC curve.

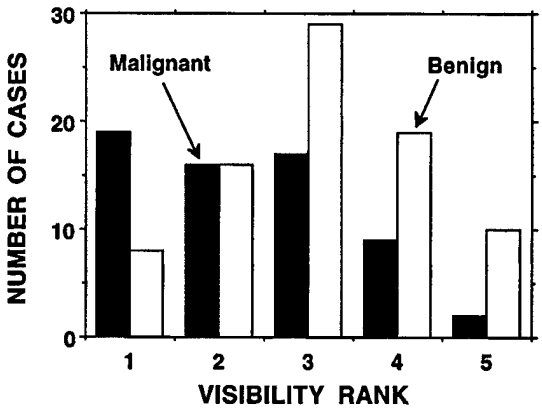


FIG. 1. Distribution of the visibility rankings of the 145 clusters of microcalcifications. Higher ranking corresponds to more subtle clusters.

II. MATERIALS AND METHODS

A. Data set

The data set for this study consisted of 145 clusters of microcalcifications from mammograms of 78 patients. The cases were selected from the patient files in the Department of Radiology at the University of Michigan. The only selection criterion was that it included a biopsy-proven microcalcification cluster. We kept the number of malignant and benign cases reasonably balanced so that 82 benign and 63 malignant clusters were included. All mammograms were acquired with a contact technique using mammography systems accredited by the American College of Radiology (ACR). The dedicated mammographic systems had molybdenum anode and molybdenum filter, 0.3 mm nominal focal spot, reciprocating grid, and Kodak MinR/MinR E screen-film systems with extended processing. A radiologist experienced in mammography ranked the visibility of each microcalcification cluster on a scale of 1 (obvious) to 5 (subtle), relative to the visibility range of microcalcification clusters encountered in clinical practice. The histogram of the visibility ranking of the 145 clusters is shown in Fig. 1. The histogram indicated the mix of subtle and obvious clusters included in the data set.

The selected mammograms were digitized with a laser scanner (Lumisys DIS-1000) at a pixel size of 0.035 mm $\times$0.035 mm and 12-bit gray levels. The digitizer has an optical density (O.D.) range of about 0 to 3.5. The O.D. on the film was digitized linearly to pixel value at a calibration of 0.001 O.D. unit/pixel value in the O.D. range of about 0 to 2.8. The digitizer deviated from a linear response at O.D. higher than 2.8.

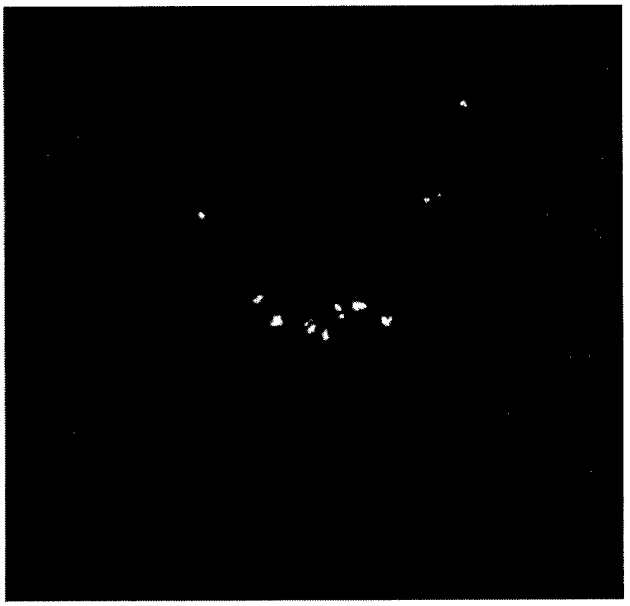
B. Morphological feature space

For the extraction of morphological features, the locations of the individual microcalcifications have to be known. We have developed an automated program for detection of individual microcalcifications.[38] However, the detection sensitivity is not 100% and the detected signals include false-positives. Furthermore, automated detection tends to have a higher likelihood of detecting obvious microcalcifications than subtle ones, which may bias the evaluation of the classification capability of the extracted features and the trained classifiers if microcalcifications detected by the automated program are used for classifier development. Since these variables are program dependent, we isolated the detection problem from the classification problem in this study by using manually identified true microcalcifications for the morphological feature analysis. The true microcalcifications were defined as those visible on the film mammograms with a magnifier. Magnification mammograms were used occasionally for verification when they were available, but in most cases only contact mammograms were used. At present, there is no other method that can more reliably identify individual microcalcifications on mammograms. Specimen radiographs can confirm the presence of the microcalcifications but the locations of the individual microcalcifications cannot be correlated with those on the mammograms because of the very different imaging geometry and techniques.

We have developed an automated signal extraction program to determine the size, contrast, signal-to-noise ratio (SNR), and shape of the microcalcifications from a mammogram based on the coordinate of each individual microcalcification. In a local region of 101×101 pixels centered at each signal site, the low frequency structured background is estimated by polynomial curve fitting in the horizontal and vertical directions and then averaging the fitted values obtained in the two directions at each pixel. This background estimation method is used because it can approximate the background more closely than two-dimensional surface fitting or the distance-weighted interpolation method (described below) used for texture feature extraction. The central $l\times l$ pixels that contain the signal are excluded from the curve fitting and noise estimation. The size l is chosen to be a constant of 15 pixels which is larger than the diameters of the microcalcifications of interest yet much smaller than the local region. The background pixel values in this $l\times l$ region are estimated from the fitted and smoothed background surface. The exclusion of the signal region is necessary so that the high contrast pixel values of the microcalcification will not affect the background estimation at the signal site. Other microcalcifications that may locate within the 101×101 pixel region are treated as background pixels because their effect on the estimated background levels at the signal site will be relatively small.

After subtraction of the structured background, the local root-mean-square (rms) noise is calculated. A gray-level threshold is determined as the product of the rms noise and an input SNR threshold. With a region growing technique, the signal region is then extracted as the connected pixels above the threshold around the manually identified signal location. A high threshold will result in extracting only the peak pixels of the microcalcification which may not represent its shape perceived on the mammogram. A low threshold will cause the microcalcification region to grow into the surrounding background pixels. Since there is no objective standard what the actual shape of a microcalcification is on a mammogram, the proper threshold to extract the signals was

(a)



(b)

FIG. 2. An example of a cluster of malignant microcalcifications in the data set: (a) the cluster with mammographic background, (b) the cluster after segmentation. Morphological features are extracted from the segmented microcalcifications.

determined by visually comparing the microcalcifications in the original image and the thresholded image of the microcalcifications superimposed on a background of constant pixel values. After an experienced radiologist compared a subset of randomly selected microcalcification clusters extracted at different thresholds, an SNR threshold of 2.0 was chosen for all cases. An example of a malignant cluster and the microcalcifications extracted at an SNR threshold of 2.0 is shown in Fig. 2.

The feature descriptors determined from the extracted microcalcifications are listed in Table I. The size of a microcalcification (SA) is estimated as the number of pixels in the signal region. The mean density (MD) is the average of the pixel values above the background level within the signal region. The second moments are calculated as

TABLE I. The 21 morphological features extracted from a microcalcification cluster.

	Average	Standard deviation	Coefficient of variation	Maximum
Area	AVSA	SDSA	CVSA	MXSA
Mean density	AVMD	SDMD	CVMD	MXMD
Eccentricity	AVEC	SDEC	CVEC	MXEC
Moment ratio	AVMR	SDMR	CVMR	MXMR
Axis ratio	AVAR	SDAR	CVAR	MXAR
No. of microcalcifications in cluster	NUMS			

$$M_{xx}=\sum_i g_i(x_i-M_x)^2/M_0, \tag{1}$$

$$M_{yy}=\sum_i g_i(y_i-M_y)^2/M_0, \tag{2}$$

$$M_{xy}=\sum_i g_i(x_i-M_x)(y_i-M_y)/M_0, \tag{3}$$

where g_i is the pixel value above the background, and (x_i,y_i) are the coordinates of the ith pixel. The moments M_0, M_x and M_y are defined as follows:

$$M_0=\sum_i g_i, \tag{4}$$

$$M_x=\sum_i g_i x_i/M_0, \tag{5}$$

$$M_y=\sum_i g_i y_i/M_0. \tag{6}$$

The summations are over all pixels within the signal region. The lengths of the major axis, $2a$, and the minor axis, $2b$, of the effective ellipse that characterizes the second moments are given by

$$2a=\sqrt{2[M_{xx}+M_{yy}+\sqrt{(M_{xx}-M_{yy})^2+4M_{xy}^2}]}, \tag{7}$$

$$2b=\sqrt{2[M_{xx}+M_{yy}-\sqrt{(M_{xx}-M_{yy})^2+4M_{xy}^2}]}. \tag{8}$$

The eccentricity (EC) of the effective ellipse can be derived from the major and minor axes as

$$\epsilon=\frac{\sqrt{a^2-b^2}}{a}. \tag{9}$$

The moment ratio (MR) is defined as the ratio of M_{xx} to M_{yy}, with the larger second moment in the denominator. The axis ratio (AR) is the ratio of the major axis to the minor axis of the effective eclipse.

To quantify the variation of the visibility and shape descriptors in a cluster, the maximum (MX), the average (AV) and the standard deviation (SD) of each feature for the individual microcalcifications in the cluster are calculated. The coefficient of variation (CV), which is the ratio of the SD to AV, is used as a descriptor of the variability of a certain

feature within a cluster. Twenty cluster features are therefore derived from the five features (size, mean density, moment ratio, axis ratio, and eccentricity) of the individual microcalcifications. Another feature describing the number of microcalcifications in a cluster (NUMS) is also added, resulting in a 21-dimensional morphological feature space.

C. Texture feature space

Our texture feature extraction method has been described in detail previously.[31] Briefly, texture features are extracted from a 1024×1024 pixel region of interest (ROI) that contains the cluster of microcalcifications. Most of the clusters in this data set can be contained within the ROI. For the few clusters that are substantially larger than a single ROI, additional ROIs containing the remaining parts of the cluster are extracted and processed in the same way as the other ROIs. The texture feature values extracted from the different ROIs of the same cluster are averaged and the average values are used as the feature values for that cluster.

For a given ROI, background correction is first performed to reduce the low frequency gray-level variation due to the density of the overlapping breast tissue and the x-ray exposure conditions. The gray level at a given pixel of the low frequency background is estimated as the average of the distance-weighted gray levels of four pixels at the intersections of the normals from the given pixel to the four edges of the ROI.[39] The estimated background image was subtracted from the original ROI to obtain a background-corrected image. An example of the background correction procedure is shown in Fig. 3.

As discussed in our previous study,[31] it was found that the texture features derived from the SGLD matrix of the ROI provided useful texture information for classification of microcalcification clusters. The SGLD matrix element, $p_{\theta,d}(i,j)$, is the joint probability of the occurrence of gray levels i and j for pixel pairs which are separated by a distance d and at a direction θ.[40] The SGLD matrices were constructed from the pixel pairs in a subregion of 512×512 pixels centered approximately at the center of the cluster in the background-corrected ROI so that any potential edge effects caused by background correction will not affect the texture extraction. We analyzed the texture features in four directions: $\theta=0°$, 45°, 90°, and 135° at each pixel pair distance d. The pixel pair distance was varied from 4 to 40 pixels in increments of 4 pixels. Therefore, a total of 40 SGLD matrices were derived from each ROI. The SGLD matrix depends on the bin width (or gray-level interval) used in accumulating the histogram. Based on our previous study, a bin width of four gray levels was chosen for constructing the SGLD matrices. This is equivalent to reducing the gray-level resolution (or bit depth) of the 12-bit image to 10 bits by eliminating the 2 least significant bits.

From each of the SGLD matrices, we derived 13 texture measures including correlation, entropy, energy (angular second moment), inertia, inverse difference moment, sum average, sum entropy, sum variance, difference average, difference entropy, difference variance, information measure of correlation 1, and information measure of correlation 2. The formulation of these texture measures could be found in the literature.[31,40] As found in our previous study,[41] we did not observe a significant dependence of the discriminatory power of the texture features on the direction of the pixel pairs for mammographic textures. However, since the actual distance between the pixel pairs in the diagonal direction was a factor

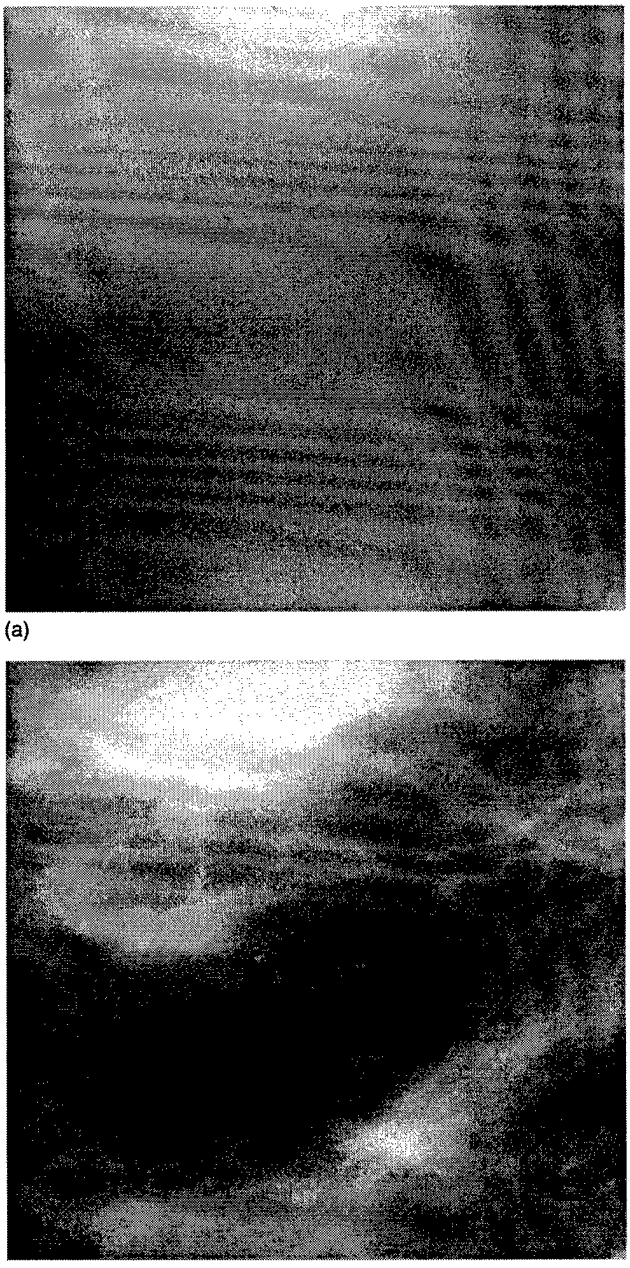


FIG. 3. An example of background correction for the ROIs before texture feature extraction. The ROI from the original image is shown in Fig. 2(a). (a) The estimated low frequency background gray level, and (b) the ROI after background correction. The background gray-level variation due to the varying x-ray penetration in the breast tissue is reduced. The contouring in the background image is a display artifact that does not exist in the calculated image file. For display purpose, the background-corrected ROI is contrast-enhanced to improve the visibility of the microcalcifications and the detailed structures.

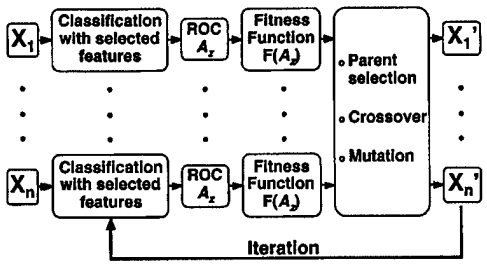


FIG. 4. A schematic diagram of the genetic algorithm designed for feature selection used in this study. $X_1,\ldots,X_n$ represents the set of parent chromosomes and $X_1',\ldots,X_n'$ represents the set of offspring chromosomes.

of $\sqrt{2}$ greater than that in the axial direction, we averaged the feature values in the axial directions (0° and 90°) and in the diagonal directions (45° and 135°) separately for each texture feature derived from the SGLD matrix at a given pixel pair distance. The average texture features at the ten pixel pair distances and two directions formed a 260-dimensional texture feature space.

D. Feature selection

Feature selection is one of the most important steps in classifier design because the presence of ineffective features often degrades the performance of a classifier on test samples. This is partly caused by the "curse of dimensionality" problem that the classifier is inadequately trained in a large-dimension feature space when only a finite number of training samples is available.[42–45] We compared two feature selection methods to extract useful features from the morphological, texture, and the combined feature spaces. One is a genetic algorithm approach, and the other is the commonly used stepwise linear discriminant analysis method.

1. Genetic algorithm for feature selection

The genetic algorithm (GA) methodology was first introduced by Holland in the early 1970s.[32,33] A GA solves an optimization problem based on the principles of natural selection. In natural selection, a population evolves by finding beneficial adaptations to a complex environment. The characteristics of a population are carried onto the next generation by its chromosomes. New characteristics are introduced into a chromosome by crossover and mutation. The probability of survival or reproduction of an individual depends more or less on its fitness to the environment. The population therefore evolves toward better-fit individuals.

The application of GA to feature selection has been described in the literature.[46,47] We have demonstrated previously that a GA could select effective features for classification of masses and normal breast tissue from a very large-dimension feature space.[34] The GA was adapted to the current problem for classification of malignant and benign microcalcifications. A brief outline is given as follows. Each feature in a given feature space is treated as a gene and is encoded by a binary digit (bit) in a chromosome. A "1" represents the presence of the feature and a "0" represents the absence of the feature. The number of genes (bits) on a chromosome is equal to the dimensionality (k) of the feature space, but only the features that are encoded as "1" are actually present in the subset of selected features. A chromosome therefore represents a possible solution to the feature selection problem.

The implementation of GA for feature selection is illustrated in the block diagram shown in Fig. 4. To allow for diversity, a large number, n, of chromosomes, $X_1,\ldots, X_n$, is chosen as the population. The number of chromosomes is kept constant in each generation. At the initiation of the GA, each bit on a chromosome is initialized randomly with a small but equal probability, P_{init}, to be "1." The selected feature subset on a chromosome is used as the input feature variables to a classifier, which was chosen to be the Fischer's linear discriminant in this study.

The available samples in the dataset are randomly partitioned into a training set and a test set. The training set is used to formulate a linear discriminant function with each of the selected feature subsets. The effectiveness of each of the linear discriminants for classification is evaluated with the test set. The classification accuracy is determined as the area, A_z, under the ROC curve. To reduce biases in the classifiers due to case selection, training and testing are performed a large number of times, each with a different random partitioning of the data set. In this study, we chose to partition the dataset 80 times and the 80 test A_z values were averaged and used for determination of the fitness of the chromosome.

The fitness function for the ith chromosome, $F(i)$, is formulated as

$$F(i)=\left[\frac{f(i)-f_{\min}}{f_{\max}-f_{\min}}\right]^2, \quad i=1,\ldots,n, \tag{10}$$

where

$$f(i)=\overline{A_z(i)}-\alpha N(i),$$

$\overline{A_z(i)}$ is the average test A_z for the ith chromosome over the 80 random partitions of the data set, $f_{\min}$ and $f_{\max}$ are the minimum and maximum $f(i)$ among the n chromosomes, $N(i)$ is the number of features in the ith chromosome, and α is a penalty factor, whose magnitude is less than $1/k$, to suppress chromosomes with a large number of selected features. The value of the fitness function $F(i)$ ranges from 0 to 1. The probability of the ith chromosome being selected as a parent, $P_s(i)$, is proportional to its fitness function:

$$P_s(i)=F(i)/\sum_{i=1}^{n} F(i), \quad i=1,\ldots,n. \tag{11}$$

A random sampling based on the probabilities, $P_s(i)$, will allow chromosomes with higher value of fitness to be selected more frequently.

For every pair of selected parent chromosomes, X_i and X_j, a random decision is made to determine if crossover should take place. A uniform random number in (0,1] is generated. If the random number is greater than P_c, the probability of crossover, then no crossover will occur; otherwise, a random crossover site is selected on the pair of chromosomes. Each chromosome is split into two strings at this site and one of the strings will be exchanged with the corre-

sponding string from the other chromosome. Crossover results in two new chromosomes of the same length.

After crossover, another chance of introducing new features is obtained by mutation. Mutation is applied to each gene on every chromosome. For each bit, a uniform random number in (0,1] is generated. If the random number is greater than P_m, the probability of mutation, then no mutation will occur; otherwise, the bit is complemented. The processes of parent selection, crossover, and mutation result in a new generation of n chromosomes, $X_1', \ldots, X_n'$, which will again be evaluated with the 80 training and test set partitions as described above. The chromosomes are allowed to evolve over a preselected number of generations. The best subset of features is chosen to be the chromosome that provides the highest average A_z during the evolution process.

In this study, 500 chromosomes were used in the population. Each chromosome has 281 gene locations. P_{init} was chosen to be 0.01 so that each chromosome started with two to three features on the average. We varied P_c from 0.7 to 0.9, P_m from 0.001 to 0.005, and α from 0 to 0.001. These ranges of parameters were chosen based on our previous experience with other feature selection problems using GA.[34]

2. Stepwise linear discriminant analysis

The stepwise linear discriminant analysis (LDA) is a commonly used method for selection of useful feature variables from a large feature space. Detailed descriptions of this method can be found in the literature.[35] The procedure is briefly outlined below. The stepwise LDA uses a forward selection and backward removal strategy. When a feature is entered into or removed from the model, its effect on the separation of the two classes can be analyzed by several criteria. We use the Wilks' lambda criterion which minimizes the ratio of the within-group sum of squares to the total sum of squares of the two class distributions; the significance of the change in the Wilks' lambda is estimated by F-statistics. In the forward selection step, the features are entered one at a time. The feature variable that causes the most significant change in the Wilks' lambda will be included in the feature set if its F value is greater than the F-to-enter (F_{in}) threshold. In the feature removal step, the features already in the model are eliminated one at a time. The feature variable that causes the least significant change in the Wilks' lambda will be excluded from the feature set if its F value is below the F-to-remove (F_{out}) threshold. The stepwise procedure terminates when the F values for all features not in the model are smaller than the F_{in} threshold and the F values for all features in the model are greater than the F_{out} threshold. The number of selected features will decrease if either the F_{in} threshold or the F_{out} threshold is increased. Therefore, the number of features to be selected can be adjusted by varying the F_{in} and F_{out} values.

E. Classifier

The training and testing procedure described above was used for the purpose of feature selection only. After the best subset of features as determined by either the GA or the stepwise LDA procedure was found, we performed the classification as follows.

The linear discriminant analysis[36] procedure in the SPSS software package[35] was used to classify the malignant and benign microcalcification clusters. We used a cross-validation resampling scheme for training and testing the classifier. The data set of 145 samples was randomly partitioned into a training set and a test set by an approximately 3:1 ratio. The partitioning was constrained so that ROIs from the same patient were always grouped into the same set. The training set was used to determine the coefficients (or weights) of the feature variables in the linear discriminant function. The performance of the trained classifier was evaluated with the test set. In order to reduce the effect of case selection, the random partitioning was performed 50 times. The results were then averaged over the 50 partitions.

The classification accuracy of the LDA was evaluated by ROC methodology. The output discriminant score from the LDA classifier was used as the decision variable in the ROC analysis. The LABROC program,[37] which assumes binormal distributions of the decision variable for the two classes and fits an ROC curve to the classifier output based on maximum-likelihood estimation, was used to estimate the ROC curve of the classifier. The ROC curve represents the relationship between the true-positive fraction (TPF) and the false-positive fraction (FPF) as the decision threshold varies. The area under the ROC curve and the standard deviation of the A_z were provided by the LABROC program for each partition of training and test sets. The average performance of the classifier was estimated as the average of the 50 test A_z values from the 50 random partitions.

To obtain a single distribution of the discriminant scores for the test samples, we performed a leave-one-case-out resampling scheme for training and testing the classifier. In this scheme, one of the 78 cases was left out at a time and the clusters from the other 77 cases were used for formulation of the linear discriminant function. The resulting LDA classifier was used to classify the clusters from the left-out case. The procedure was performed 78 times so that every case was left out once to be the test case. The test discriminant scores from all the clusters were accumulated in a distribution which was then analyzed by the LABROC program. Using the distributions of discriminant scores for the test samples from the leave-one-case-out resampling scheme, the CLABROC program could be used to test the statistical significance of the differences between ROC curves[48] obtained from different conditions. The two-tailed p value for the difference in the areas under the ROC curves was estimated.

III. RESULTS

The variations of best feature set size and classifier performance in terms of A_z with the GA parameters were tabulated in Table II(a)–(c) for the morphological, the texture, and the combined feature spaces, respectively. The number of generations that the chromosomes evolved was fixed at 75

TABLE II. Dependence of feature selection and classifier performance on GA parameters: (a) morphological feature space, (b) texture feature space, and (c) combined feature space. The number of generations that the GA evolved was fixed at 75. The best result for each feature space is identified with an asterisk.

(a)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	6	0.84	0.79
0.8			3	0.77	0.76
0.9			4	0.80	0.77
0.7	0.003		7	0.82	0.78
0.8			6	0.82	0.79
0.9			6	0.84	0.79
0.7	0.001	0.0005	3	0.77	0.76
0.8			4	0.80	0.77
0.9			3	0.77	0.76
0.7	0.003		6	0.84	0.79*
0.8			6	0.84	0.79
0.9			6	0.82	0.79
0.7	0.001	0.0010	3	0.77	0.76
0.8			4	0.80	0.77
0.9			3	0.77	0.76
0.7	0.003		6	0.84	0.79
0.8			7	0.84	0.79
0.9			4	0.80	0.77

(b)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	7	0.87	0.82
0.8			8	0.88	0.84
0.9			8	0.88	0.84
0.7	0.003		17	0.91	0.82
0.8			9	0.88	0.79
0.9			10	0.88	0.79
0.7	0.001	0.0005	9	0.88	0.85*
0.8			7	0.86	0.82
0.9			8	0.87	0.84
0.7	0.003		13	0.90	0.81
0.8			10	0.87	0.81
0.9			12	0.88	0.81
0.7	0.001	0.0010	7	0.87	0.83
0.8			9	0.88	0.83
0.9			8	0.88	0.83
0.7	0.003		10	0.88	0.83
0.8			21	0.94	0.82
0.9			12	0.88	0.80

(c)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	13	0.93	0.88
0.8			12	0.92	0.88
0.9			12	0.92	0.89
0.7	0.003		12	0.91	0.86
0.8			16	0.94	0.88
0.9			17	0.95	0.88
0.7	0.001	0.0003	12	0.92	0.87
0.8			12	0.92	0.86
0.9			12	0.93	0.88
0.7	0.003		13	0.93	0.87
0.8			13	0.93	0.88
0.9			12	0.94	0.89*
0.7	0.005		12	0.89	0.80
0.7	0.001	0.0010	11	0.92	0.87
0.8			10	0.91	0.87
0.9			11	0.91	0.86
0.7	0.003		10	0.91	0.86
0.8			14	0.93	0.87
0.9			13	0.92	0.87
0.7	0.005		11	0.89	0.81
0.8			12	0.88	0.82
0.9			12	0.89	0.81

TABLE III. Dependence of feature selection and classifier performance on F_{out} and F_{in} thresholds using stepwise linear discriminant analysis: (a) morphological feature space, (b) texture feature space, and (c) combined feature space. The best result for each feature space is identified with an asterisk. When the test A_z is comparable, the feature set with fewer number of features is considered to be better.

(a)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
2.7	3.8	2	0.76	0.76
1.7	2.8	4	0.79	0.76
1.7	1.8	6	0.83	0.79*
1.0	1.4			
1.0	1.2	7	0.84	0.79
0.8	1.0	9	0.85	0.79
0.6	0.8			
0.4	0.6	10	0.85	0.79
0.2	0.4	12	0.86	0.78
0.1	0.2			

(b)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
2.7	3.8	4	0.82	0.80
1.7	2.8			
1.0	1.4	8	0.88	0.83
1.0	1.2	10	0.89	0.82
0.8	1.0	11	0.89	0.83
0.6	0.8	14	0.91	0.85*
0.4	0.6	17	0.92	0.84
0.2	0.4	18	0.92	0.81
0.1	0.2	16	0.90	0.80

(c)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
3.0	3.2	6	0.84	0.80
2.9	3.2			
2.8	3.1			
2.0	3.1			
3.0	3.1	10	0.88	0.83
2.9	3.0			
2.7	2.8			
2.0	2.3	11	0.90	0.86
2.0	2.2			
1.9	2.0			
1.7	1.8			
1.3	1.5	14	0.92	0.86
1.0	1.2	19	0.95	0.86
1.0	1.1	23	0.96	0.87*
0.8	1.2	28	0.97	0.86

in these tables. The training and test A_z values were obtained from averaging results of the 50 partitions of the data sets using the selected feature sets.

The results of feature selection using the stepwise LDA procedure with a range of F_{in} and F_{out} thresholds were tabulated in Table III(a)–(c). The thresholds were varied so that the number of selected features varied over a wide range. Often different choices of F_{in} and F_{out} values could result in the same selected feature set as shown in the tables by the number of features in the set. The average A_z values obtained from the 50 partitions of the data set using the selected feature sets were listed. The best feature sets selected in the different feature spaces are shown in Table IV.

TABLE IV. The best feature sets selected by the GA and stepwise LDA methods (indicated by asterisk in Tables II and III) in the three feature spaces. The number of generations for chromosome evolution in the GA algorithm to reach the selected feature sets is listed. The abbreviations for the texture features are: correlation (CORE), energy (ENER), entropy (ENTR), difference average (DFAV), difference entropy (DFEN), difference variance (DFVR), inertia (INER), inverse difference moment (INVD), information measure of correlation 1 (ICO1), information measure of correlation 2 (ICO2), sum average (SMAV), sum entropy (SMEN), sum variance (SMVR). After an abbreviation, the letter "A" indicates diagonal features and the number indicates the pixel distance. The abbreviations for the morphological features can be found in Table I.

GA			Stepwise LDA		
Morphological generation 39	Texture generation 64	Combined generation 169	Morphological	Texture	Combined
CMVD	DFAVA_8	DFAVA_4	AVMD	DFAV_12	CORE_40
CVMR	DFEN_16	DFEN_28	CVMD	DFEN_4	COREA_16
CVSA	DFVRA_24	DFVRA_36	CVMR	DFEN_8	COREA_40
MXMR	DFVR_24	DFVR_12	CVSA	DFENA_12	DFAVA_8
MXSA	DFVR_4	DFVR_20	MXMR	DFENA_24	DFEN_4
SDMD	DFVR_8	ICO1A_20	MXSA	DFVR_24	DFEN_8
	ICO1A_12	ICO1A_32		DFVR_40	DFENA_36
	ICO2A_28	SMEN_16		ICO1_16	DFVR_20
	ICO2_40	SMEN_36		ICO1A_8	ICO1A_28
		AVAR		ICO2_40	ICO2_24
		CVMD		INER_8	ICO2_36
		CVSA		INVD_16	INER_12
		MXEC		INVD_4	INERA_16
		NUMS		INVDA_8	INVDA_36
		SDMD			SMEN_40
					SMENA_4
					AVAR
					CVMD
					CVSA
					MXAR
					MXEC
					NUMS
					SDMD

Table V compares the training and test A_z values from the best feature set in each feature space for the two feature selection methods. The GA parameters that selected the feature set with best classification performance in each feature space after 75 generations (Table II) were used to run the GA again for 500 generations. The A_z values obtained with the best GA selected feature sets after 75 generations are listed together with those obtained after 500 generations. The A_z values obtained with the leave-one-case-out scheme are also shown in Table V. The differences between the corresponding A_z values from the two resampling schemes are within 0.01. The two feature selection methods provided feature sets that had similar test A_z values in the morphological and texture feature spaces. In the combined feature space, there was a slight improvement in the test A_z value obtained with the GA selected features. Although the difference in the A_z

TABLE V. Classification accuracy of linear discriminant classifier in the different feature spaces using feature sets selected by the GA and the stepwise LDA procedure.

	Training A_z			Text A_z		
Feature selection	Morphological	Texture	Combined	Morphological	Texture	Combined
Cross-validation						
GA (75 generations)	0.84±0.04	0.88±0.03	0.94±0.02	0.79±0.07	0.85±0.07	0.89±0.05
GA (500 generations)	0.84±0.04	0.88±0.03	0.96±0.02	0.79±0.07	0.85±0.07	0.90±0.05
Stepwise LDA	0.83±0.04	0.91±0.03	0.96±0.02	0.79±0.07	0.85±0.06	0.87±0.06
Leave-one-case-out						
GA (75 generations)	0.83±0.03	0.88±0.03	0.94±0.02	0.79±0.04	0.84±0.03	0.89±0.03
GA (500 generations)	0.83±0.03	0.88±0.03	0.95±0.02	0.79±0.04	0.84±0.03	0.89±0.03
Stepwise LDA	0.83±0.03	0.91±0.02	0.96±0.02	0.79±0.04	0.85±0.03	0.87±0.03

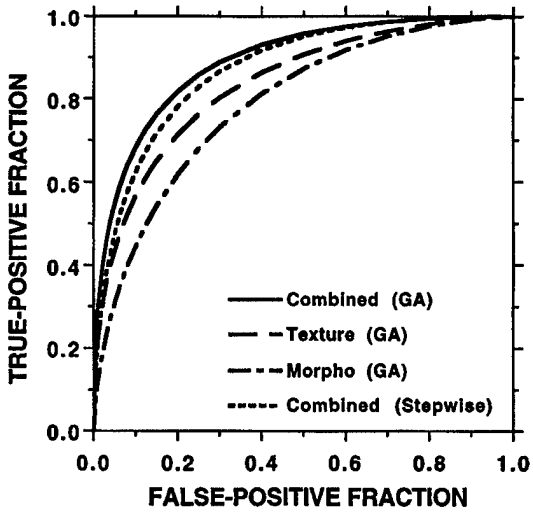


FIG. 5. Comparison of ROC curves of the LDA classifier performance using the best GA selected feature sets in the three feature spaces. In addition, the ROC curve obtained from the best feature set selected by the stepwise LDA procedure in the combined feature space is shown. The classification was performed with a leave-one-case-out resampling scheme.

values from the leave-one-case-out scheme between the two feature selection methods did not achieve statistical significance ($p=0.2$), as estimated by CLABROC, the differences in the paired A_z values from the 50 partitions demonstrated a consistent trend (40 out of 50 partitions) that the A_z from the GA selected features were higher than those obtained by the stepwise LDA. This trend was also observed in our previous study in which mass and normal tissue were classified.[34]

The ROC curves for the test samples using the feature sets selected by the GA were plotted in Fig. 5. The classification accuracy in the combined feature space was significantly higher than those in the morphological ($p=0.002$) or the texture feature space ($p=0.04$) alone. The ROC curve using the feature set selected by the stepwise procedure in the combined feature space was also plotted for comparison. The distribution of the discriminant scores for the test samples using the feature set selected by the GA in the combined feature space is shown in Fig. 6(a). If a decision threshold is chosen at 0.3, 29 of the 82 (35%) benign samples can be correctly classified without missing any malignant clusters.

Some of the 145 samples are different views of the same microcalcification clusters. In clinical practice, the decision regarding a cluster is based on information from all views. If it is desirable to provide the radiologist a single relative malignancy rating for each cluster, two possible strategies may be used to merge the scores from all views: the average score or the minimum score. The latter strategy corresponds to the use of the highest likelihood of malignancy score for the cluster. There were a total of 81 different clusters (44 benign and 37 malignant) from the 78 cases because 3 of the cases contained both a benign and a malignant cluster. The distributions of the average and the minimum discriminant scores of the 81 clusters in the combined feature space were plotted in Fig. 6(b) and Fig. 6(c), respectively. Using the average scores, ROC analysis provided test A_z values of 0.93 ± 0.03 and 0.89 ± 0.04, respectively, for the GA selected and stepwise LDA selected feature sets. Using the minimum scores, the test A_z values were 0.90 ± 0.03 and 0.85 ± 0.04, respectively. The difference between the A_z values from the two feature selection methods did not achieve statistical significance in either case ($p=0.07$ and $p=0.09$, respectively). If a decision threshold is chosen at an average score of 0.2, 22 of the 44 (50%) benign clusters can be correctly identified with 100% correct classification of the malignant clusters. If a decision threshold is set at a minimum score of 0.2, 14 of the

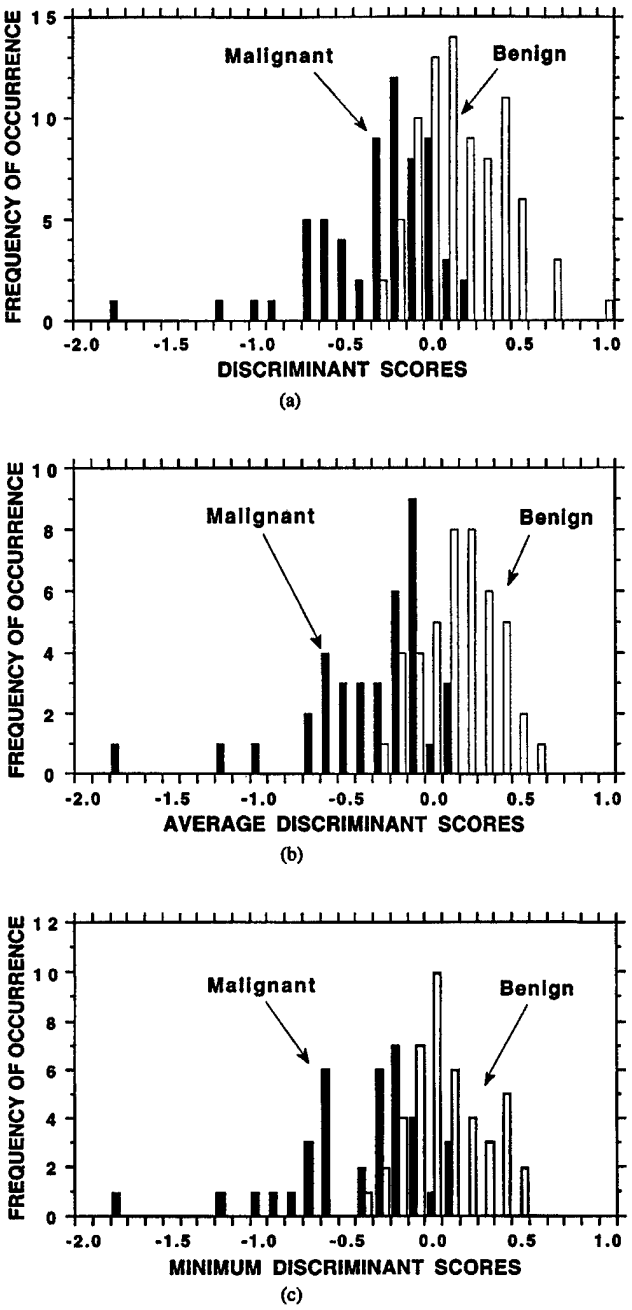


FIG. 6. Distribution of the discriminant scores for the test samples using the best GA selected feature set in the combined texture and morphological feature space. (a) Classification by samples from each film, (b) classification by cluster using the average scores, (c) classification by cluster using the minimum scores.

44 (32%) benign clusters can be identified at 100% sensitivity.

IV. DISCUSSION

The Fischer's linear discriminant is the optimal classifier if the class distributions are multivariate normal with equal covariance matrices.[42] Even if these conditions are not satisfied, as in most classification tasks, the LDA may still be a preferred choice when the number of available training samples is small. Our previous investigation[43,45] of the dependence of classifier performance on design sample size indicated that, in general, the training performance (resubstitution) of a classifier is positively biased whereas the test performance (hold-out) is negatively biased by the sample size. The magnitudes of the biases increase when the dimensionality of the input feature space or the complexity of the classifier increases, or when the design sample size decreases. Therefore, the test performance of a linear classifier is generally better than that of a more complex classifier such as a neural network or a quadratic classifier when the training sample size is small. The training results should not be used for comparison of classifier performance because a classifier can often be overtrained and give a near-perfect classification on training samples while the generalization to any unknown test samples is poor. In this study, we evaluated the effectiveness of using the morphological and the texture features extracted from mammograms for classification of a microcalcification cluster. Although we expanded the data set from our previous study, the current data set was still relatively small. We therefore chose to use a linear discriminant classifier for this classification task. Stepwise feature selection or a GA was used to reduce the dimensionality of the feature space.

In the morphological feature space, the features related to three characteristics, mean density, the moment ratio, and the signal area, were chosen most often. The features related to axis ratio, eccentricity, and the number of microcalcifications in a cluster were chosen only when they were combined with texture features. These results indicate the usefulness of classification in multi-dimensional feature spaces. Some features that are not useful by themselves can become effective features when they are combined with other features. The results also indicate that all six characteristics of the microcalcifications designed for this task have some discriminatory power to distinguish malignant and benign microcalcifications. The morphological features are not as effective as the texture features. This is evident from the smaller A_z values in the morphological feature space. However, when the morphological feature space is combined with the texture feature space, the resulting feature set selected from the combined feature space can significantly improve the classification accuracy, in comparison with those from the individual feature spaces.

The SGLD texture features characterize the shape of the SGLD matrix and generally contain information about the image properties such as homogeneity, contrast, the presence of organized structures, as well as the complexity and gray-level transitions within the image.[40] As an example, the entropy feature measures the uniformity of the SGLD matrix. The entropy value is maximum when all the matrix elements are equal. The entropy value is small when large matrix elements concentrate in a small region of the SGLD matrix while the other matrix elements are relatively small. Therefore, large entropy represents a large but random variation of pixel values in an image without regular structures whereas small entropy represents an image with relatively uniform pixel values if the SGLD matrix peaks along the diagonal and an image with regular texture patterns if it peaks off the diagonal. The ambiguity may be resolved when the sum entropy and difference entropy measures are analyzed. Unlike morphological features, it is difficult, in general, to find the direct relationship between a texture measure and the structures seen on an image,[40] and often a combination of several texture measures extracted at different angles and pixel pair distances are required to describe a texture pattern. It may also be noted that some textures can only be described by second-order statistics and may not be distinguishable by human eyes. The feature selection methods are used to empirically find the combination of features that can most effectively distinguish the malignant and benign lesions.

From Table IV, it can be seen that many of the features in the best feature sets selected by the GA method and the stepwise LDA method are similar. In the morphological feature space, five of the six selected features are the same in the two feature sets. In the combined feature space, six morphological features (out of six and seven morphological features in the two sets, respectively) are the same. For the texture features, there are more variations in the features selected by the two methods. However, the differences are mainly in the pixel distances and the directions of the features, while the major types of the texture features are similar. For example, four types of texture features, energy, entropy, sum average, and sum variance were not selected in either the texture or the combined feature space by both methods. Another four types of texture features, difference average, difference entropy, difference variance, and information measure of correlation 1 were chosen in each case, and information measure of correlation 2 was chosen in three of the four cases. Inertia and inverse difference moment were selected by the stepwise LDA method in both the texture and the combined feature spaces. Sum entropy was selected by both methods in the combined feature space. These results indicate that some features are more effective than the others for distinguishing benign and malignant microcalcifications. The pixel distance and the direction of the texture features may be considered to be higher order effects that have less influence on the discriminatory ability of a given type of texture measure. The smaller differences in their discriminatory ability would subject them to greater variability of being chosen in the feature selection processes. It may also be noted that many of the features are highly correlated. The correlated features can be interchanged in a classifier model without a strong effect on its performance.

The GA solves an optimization problem based on a search guided by the fitness function. Ideally, the values for the P_m,

P_c, and α parameters chosen in the GA only affect the convergence rate but will eventually evolve to the same global maximum. However, when the dimensionality of the feature space is very large and the design samples are sparse, the GA often reaches local maxima corresponding to different feature sets, as can be seen in Table II. Similarly, the stepwise feature selection may reach a different local maximum and choose a feature set different from those chosen by the GA. The different feature sets may provide different or similar performance. The latter is often a result of the correlation among the features, as described above.

For the linear discriminant classifier, the stepwise LDA procedure can select near-optimal features for the classification task. We have shown that the GA could select a feature set comparable to or slightly better than that selected by the stepwise LDA. The number of generations that the GA had to evolve to reach the best selection increased with the dimensionality of the feature space as expected. However, even in a 281-dimensional feature space, it only took 169 generations to find a better feature set than that selected by stepwise LDA. Further search up to 500 generations did not find other feature combinations with better performance. Although the difference in A_z did not achieve statistical significance, probably due to the large standard deviation in A_z when the number of case samples in the ROC analysis was small, the improvements in A_z in this and our previous studies[34] indicate that the GA is a useful feature selection method for classifier design. One of the advantages of GA-based feature selection is that it can search for near-optimal feature sets for any types of linear or nonlinear classifiers, whereas the stepwise LDA procedure is more tailored to linear discriminant classifiers. Furthermore, the fitness function in the GA can be designed such that features with specific characteristics are favored. One of the applications in this direction is to select features to design a classifier with high sensitivity and high specificity for classification of malignant and benign lesions.[49,50] Although the GA requires much longer computation time than the stepwise LDA to search for the best feature set, the flexibility of the GA makes it an increasingly popular alternative for solving machine learning and optimization problems. Since feature selection is performed only during training of a classifier, the speed of a trained classifier for processing test cases is not affected by the choice of the feature selection method. Therefore, the longer computation time of GA is not a problem in practice if the GA can provide a better feature set for a given classification task.

V. CONCLUSIONS

In this study, we evaluated the effectiveness of morphological and texture features extracted from mammograms for classification of malignant and benign microcalcification clusters. We also compared a GA-based feature selection method and a stepwise feature selection procedure based on linear discriminant analysis. It was found that the best feature set was selected from the combined morphological and texture feature space by the GA-based method. A linear discriminant classifier using the best feature set and a properly chosen decision threshold could correctly identify 35% of the benign clusters without missing any malignant clusters. If the average discriminant score from all views of the same cluster was used for classification, the accuracy improved to 50% specificity at 100% sensitivity. Alternatively, if the minimum discriminant score from all views of the same cluster was used, the accuracy would be 32% specificity at 100% sensitivity. This information may be used to reduce unnecessary biopsies, thereby improving the positive predictive value of mammography. Although these results were obtained with a relatively small data set, they demonstrate the potential of using CAD techniques to analyze mammograms and to assist radiologists in making diagnostic decisions. Further studies will be conducted to evaluate the generalizability of our approach in large data sets.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 48129 and by U.S. Army Medical Research and Materiel Command Grant No. DAMD 17-96-1-6254. Berkman Sahiner is also supported by a Career Development Award by the U.S. Army Medical Research and Materiel Command (DAMD 17-96-1-6012). Nicholas Petrick is also supported by a grant from The Whitaker Foundation. The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D. for use of the LABROC and CLABROC programs.

[a]Electronic mail: chanhp@umich.edu

[1]D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," Current Opinion in Radiology **4**, 123–129 (1992).

[2]D. B. Kopans, "The positive predictive value of mammography," Am. J. Roentgenol. **158**, 521–526 (1991).

[3]M. Sabel and H. Aichinger, "Recent developments in breast imaging," Phys. Med. Biol. **41**, 315–368 (1996).

[4]F. Shtern, C. Stelling, B. Goldberg, and R. Hawkins, "Novel technologies in breast imaging: National Cancer Institute perspective," Society of Breast Imaging, Orlando, Florida, 153–156 (1995).

[5]L. V. Ackerman and E. E. Gose, "Breast lesion classification by computer and xeroradiograph," Cancer **30**, 1025–1035 (1972).

[6]J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).

[7]Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," Med. Phys. **22**, 1569–1579 (1995).

[8]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of masses on mammograms using rubber-band straightening transform and feature analysis," Proc. SPIE **2710**, 44–50 (1996).

[9]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber-band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1998).

[10]S. Pohlman, K. A. Powell, N. A. Obuchowshi, W. A. Chilote, and S. Grundfest-Broniatowski, "Quantitative classification of breast tumors in digitized mammograms," Med. Phys. **23**, 1337–1345 (1996).

[11]W. G. Wee, M. Moskowitz, N.-C. Chang, Y.-C. Ting, and S. Pemmeraju, "Evaluation of mammographic calcifications using a computer program," Radiology **116**, 717–720 (1975).

[12]S. H. Fox, U. M. Pujare, W. G. Wee, M. Moskowitz, and R. V. P. Hutter, "A computer analysis of mammographic microcalcifications: global approach.," Proceedings of the IEEE 5th International Conference on Pattern Recognition., IEEE, New York, 624–631 (1980).

[13]H. P. Chan, L. T. Niklason, D. M. Ikeda, and D. D. Adler, "Computer-aided diagnosis in mammography: Detection and characterization of microcalcifications," Med. Phys. **19**, 831 (1992).

[14]H. P. Chan, D. Wei, L. T. Niklason, M. A. Helvie, K. L. Lam, M. M. Goodsitt, and D. D. Adler, "Computer-aided classification of malignant/benign microcalcifications in mammography," Med. Phys. **21**, 875 (1994).

[15]H. P. Chan, D. Wei, K. L. Lam, S.-C. B. Lo, B. Sahiner, M. A. Helvie, and D. D. Adler, "Computerized detection and classification of microcalcifications on mammograms," Proc. SPIE **2434**, 612–620 (1995).

[16]H. P. Chan, B. Sahiner, K. L. Lam, D. Wei, M. A. Helvie, and D. D. Adler, "Classification of malignant and benign microcalcifications on mammograms using an artificial neural network," Proc. of World Congress on Neural Networks II, 889–892 (1995).

[17]H. P. Chan, D. Wei, K. L. Lam, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of malignant and benign microcalcifications by texture analysis," Med. Phys. **22**, 938 (1995).

[18]H. P. Chan, B. Sahiner, D. Wei, M. A. Helvie, D. D. Adler, and K. L. Lam, "Computer-aided diagnosis in mammography: Effect of feature classifier on characterization of microcalcifications," Radiology **197(P)**, 425 (1995).

[19]L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications," IEEE Trans. Med. Imaging **13**, 263–274 (1994).

[20]Y. Wu, M. T. Freedman, A. Hasegawa, R. A. Zuurbier, S. C. B. Lo, and S. K. Mun, "Classification of microcalcifications in radiographs of pathologic specimens for the diagnosis of breast cancer," Academic Radiology **2**, 199–204 (1995).

[21]Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," Radiology **198**, 671–678 (1996).

[22]D. L. Thiele, C. Kimme-Smith, T. D. Johnson, M. McCombs, and L. W. Bassett, "Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes," Med. Phys. **23**, 549–555 (1996).

[23]A. P. Dhawan, Y. Chitre, C. Kaiser-Bonasso, and M. Moskowitz, "Analysis of mammographic microcalcifications using gray-level image structure features," IEEE Trans. Med. Imaging **15**, 246–259 (1996).

[24]L. V. Ackerman, A. N. Mucciardi, E. E. Gose, and F. S. Alcorn, "Classification of benign and malignant breast tumors on the basis of 36 radiographic properties," Cancer **31**, 342 (1973).

[25]A. G. Gale, E. J. Roebuck, P. Riley, and B. S. Worthington, "Computer aids to mammographic diagnosis," Br. J. Radiol. **60**, 887–891 (1987).

[26]D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," Invest. Radiol. **23**, 240 (1988).

[27]C. J. D'Orsi, D. J. Getty, J. A. Swets, R. M. Pickett, S. E. Seltzer, and B. J. McNeil, "Reading and decision aids for improved accuracy and standardization of mammographic diagnosis," Radiology **184**, 619–622 (1992).

[28]Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," Radiology **187**, 81–87 (1993).

[29]J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: Prediction with artificial neural network based on BI-RADS standardization lexicon," Radiology **196**, 817–822 (1995).

[30]J. Y. Lo, J. A. Baker, P. J. Kornguth, J. D. Iglehart, and C. E. Floyd, "Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features," Radiology **203**, 159–163 (1997).

[31]H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," Phys. Med. Biol. **42**, 549–567 (1997).

[32]J. H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, MI, 1975).

[33]D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, New York, 1989).

[34]B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms," Med. Phys. **23**, 1671–1684 (1996).

[35]M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics* (SPSS Inc., Chicago, IL, 1993).

[36]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975), Chaps. 1, 3.

[37]C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously-distributed test results," Annual Meeting of the American Statistical Association, Anaheim, CA (1990).

[38]H. P. Chan, S. C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," Med. Phys. **22**, 1555–1567 (1995).

[39]B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," IEEE Trans. Med. Imaging **15**, 598–610 (1996).

[40]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).

[41]H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).

[42]K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990), Chap. 3.

[43]H. P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: Quadratic and neural network classifiers," Proc. SPIE **3034**, 1102–1113 (1997).

[44]H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis in mammography: Effects of finite sample size," Med. Phys. **24**, 1034–1035 (1997).

[45]R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Finite-sample effects and resampling plans: Applications to linear classifiers in computer-aided diagnosis," Proc. SPIE **3034**, 467–477 (1997).

[46]F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," Pattern Recognition in Practice IV, 403–413 (1994).

[47]W. Siedlecki and J. Sklansky, "A note on genetic algorithm for large-scale feature selection," Pattern Recogn. Lett. **10**, 335–347 (1989).

[48]C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance for differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging*, edited by F. Deconinck (The Hague, Martinus Nijhoff, 1984), pp. 432–445.

[49]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of malignant and benign breast masses: Development of a high-sensitivity classifier using a genetic algorithm," Radiology **201**, 256–257 (1996).

[50]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Design of a high-sensitivity classifier based on genetic algorithm: Application to computer-aided diagnosis," Phys. Med. Biol. **43**, 2853–2871 (1998).

Phys. Med. Biol. 43 (1998) 2853–2871. Printed in the UK PII: S0031-9155(98)88223-7

Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis

Berkman Sahiner†, Heang-Ping Chan, Nicholas Petrick, Mark A Helvie and Mitchell M Goodsitt

Department of Radiology, University of Michigan, Ann Arbor, USA

Received 6 October 1997

Abstract. A genetic algorithm (GA) based feature selection method was developed for the design of high-sensitivity classifiers, which were tailored to yield high sensitivity with high specificity. The fitness function of the GA was based on the receiver operating characteristic (ROC) partial area index, which is defined as the average specificity above a given sensitivity threshold. The designed GA evolved towards the selection of feature combinations which yielded high specificity in the high-sensitivity region of the ROC curve, regardless of the performance at low sensitivity. This is a desirable quality of a classifier used for breast lesion characterization, since the focus in breast lesion characterization is to diagnose correctly as many benign lesions as possible without missing malignancies. The high-sensitivity classifier, formulated as the Fisher's linear discriminant using GA-selected feature variables, was employed to classify 255 biopsy-proven mammographic masses as malignant or benign. The mammograms were digitized at a pixel size of 0.1 mm × 0.1 mm, and regions of interest (ROIs) containing the biopsied masses were extracted by an experienced radiologist. A recently developed image transformation technique, referred to as the rubber-band straightening transform, was applied to the ROIs. Texture features extracted from the spatial grey-level dependence and run-length statistics matrices of the transformed ROIs were used to distinguish malignant and benign masses. The classification accuracy of the high-sensitivity classifier was compared with that of linear discriminant analysis with stepwise feature selection (LDA_{sfs}). With proper GA training, the ROC partial area of the high-sensitivity classifier above a true-positive fraction of 0.95 was significantly larger than that of LDA_{sfs}, although the latter provided a higher total area (A_z) under the ROC curve. By setting an appropriate decision threshold, the high-sensitivity classifier and LDA_{sfs} correctly identified 61% and 34% of the benign masses respectively without missing any malignant masses. Our results show that the choice of the feature selection technique is important in computer-aided diagnosis, and that the GA may be a useful tool for designing classifiers for lesion characterization.

1. Introduction

Due to its high sensitivity, mammography is usually the first radiological examination used for the early detection of malignant breast lesions. However, the positive predictive value (PPV) of mammographic diagnosis (ratio of the number of malignancies to the total number of biopsy recommendations) is not high. Biopsies performed for mammographically suspicious non-palpable breast masses had PPVs of 20 to 30% in three studies (Hermann *et al* 1987, Hall *et al* 1988, Jacobson and Edeiken 1990). To reduce health-care costs and patient morbidity, it is desirable to increase the PPV of mammographic diagnosis

† Address for correspondence: Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, CGC B2102, Ann Arbor, MI 48109-0904, USA. E-mail address: berki@umich.edu

0031-9155/98/102853+19$19.50

while maintaining its sensitivity of cancer detection. Computerized mammographic analysis methods can potentially aid radiologists in achieving this goal.

In recent years, several researchers have developed new techniques for the classification of mammographic masses based on computer-extracted features (Brzakovic *et al* 1990, Kilday *et al* 1993, Huo *et al* 1995, Pohlman *et al* 1996, Rangayyan *et al* 1996, Sahiner *et al* 1996a, 1997, 1998). Kilday *et al* (1993) classified masses using morphological features and patient age. Brzakovic *et al* (1990) classified suspected lesions using their shape and intensity variations. Huo *et al* (1995) developed a technique to quantify the degree of spiculation of a lesion, and classified masses as malignant and benign using these spiculation measures. Pohlman *et al* (1996) developed a region growing algorithm for tumour segmentation, and used features describing the tumour shape for classification. Rangayyan *et al* (1996) used an edge acutance measure extracted from the grey-scale intensity along the normal direction to the mass shape, as well as moments to classify masses. We have developed the rubber-band straightening transform (RBST) for facilitating the extraction of effective texture features, and used the texture features extracted from the transformed image for classification (Sahiner *et al* 1996a, 1997, 1998).

A common characteristic of the above approaches is that the lesion is first segmented from the surrounding tissue, and then features are extracted from the shape and grey-level characteristics of the lesion and the surrounding tissue. The extracted features usually represent a mathematical description of characteristics that are helpful for distinguishing malignant and benign lesions. When several features are extracted for classification, it may be difficult to predict which features or feature combinations will result in more accurate classification. For example, it is known that the borders of malignant masses tend to be more irregular than those of benign masses; therefore, it is expected that the normalized radial lengths (Kilday *et al* 1993) carry useful information about the probability of malignancy of a mass. However, since the normalized radial lengths, and especially the features extracted from them (for example variance and entropy), do not exactly measure irregularity but instead merge information from a combination of border characteristics, it is difficult to predict which feature combination will yield the highest classification accuracy when used in a statistical classifier. It is known that the inclusion of inappropriate features may adversely affect classifier performance, especially when the training set is not sufficiently large (Raudys and Jain 1991, Sahiner *et al* 1996c). Therefore, in many situations, one must face the task of selecting a subset of effective features for classification.

One systematic method for feature selection is linear discriminant analysis with stepwise feature selection (LDA_{sfs}), which has been applied to feature selection problems in computer-aided diagnosis (Chan *et al* 1995, Wei *et al* 1995). LDA_{sfs} is an iterative procedure, where one feature is entered into or removed from the selected feature pool at each step by analysing its effect on a selection criterion. The nature of the stepwise selection procedure makes it imperative that the selection criterion be a statistical distance measure between the two groups to be classified. The Wilks lambda and the Mahalanobis distance are commonly used measures. Genetic algorithm (GA) based feature selection, which is capable of using any numerically computed criterion for its fitness function, is a slower but more versatile method than stepwise feature selection. We have demonstrated that when the GA fitness criterion is related to the area A_z under the receiver operating characteristic (ROC) curve, GA-based feature selection yields slightly more effective features than LDA_{sfs} (Sahiner *et al* 1996c).

In the task of lesion characterization, the cost of missing a malignancy is very high. Therefore, the performance of a classifier in the high-sensitivity (high true-positive fraction) region of the ROC curve is more important than the overall area A_z under the ROC curve. In

other words, if a classifier is to be designed for breast lesion characterization, the specificity at high levels of sensitivity is much more important than the specificity at low levels of sensitivity. Recently, Jiang *et al* (1996) developed a method for describing an ROC partial area index that may be useful as a performance measure in lesion characterization problems. Since a feature (or feature combination) that can provide a large overall A_z (or a large Wilks lambda and Mahalanobis distance) may not provide a large partial ROC area, it is important to develop a feature selection method for the design of high-sensitivity classifiers. The partial ROC area is potentially a good feature selection criterion for this application. The flexibility of a GA in the selection of its fitness function allows this index to be incorporated for feature selection.

In this study, we developed a methodology to design high-sensitivity classifiers. The design process was illustrated by the task of classifying masses on digitized mammograms as malignant or benign. A GA-based algorithm with the ROC partial area index as the feature selection criterion, in combination with Fisher's linear discriminant, was used for the design of this classifier. Texture features extracted from RBST images (Sahiner *et al* 1998) were used for classification. The performance of the high-sensitivity classifier was compared with the performance achieved by LDA_{sfs} using the Wilks lambda as the feature selection criterion.

2. Materials and methods

2.1. Data set

The mammograms used in this study were selected from the files of patients at the Radiology Department of the University of Michigan who had undergone biopsy. The mammograms were acquired with dedicated mammographic systems with 0.3 mm focal spots, molybdenum anodes, 0.03 mm thick molybdenum filters and 5:1 reciprocating grids. For recording the images, a Kodak MinR/MRE screen/film system with extended cycle processing was used. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass, and that approximately equal numbers of malignant and benign masses were present in the data set.

Our data set consisted of 255 mammograms from 104 patients. For most of the patients we had two mammograms in the data set, which were the craniocaudal and the mediolateral oblique views. However, for some of the patients, extra views such as lateral and oblique views were included in the data set. There were 128 mammograms with benign masses, of which 8 were spiculated based upon radiologist interpretation, and 127 mammograms with malignant masses, of which 62 were spiculated. Of the 104 patients evaluated in this study, 48 had malignant masses. The probability of malignancy of the biopsied mass on each mammogram was ranked by a Mammography Quality Standards Act (MQSA) approved radiologist experienced in mammographic interpretation on a scale of 1 to 10. A ranking of 1 corresponded to the masses with the most benign mammographic appearance, and a ranking of 10 corresponded to the masses with the most malignant mammographic appearance. The distribution of the malignancy ranking of the masses is shown in figure 1. The true pathology of the masses was determined by biopsy and histological analysis.

The mammograms in the data set were digitized with a Lumisys DIS-1000 laser scanner at a pixel resolution of 0.1 mm $\times$ 0.1 mm and 4096 grey levels. The digitizer was calibrated so that grey-level values were linearly proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of 0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually, with the OD range extending to 3.5.

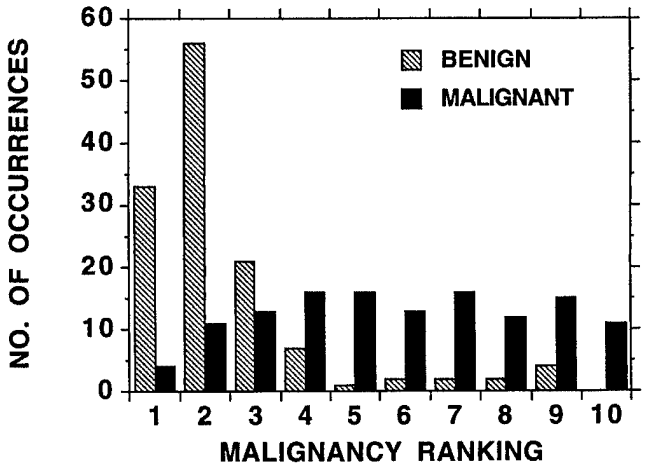


Figure 1. The distribution of the malignancy ranking of the masses in our data set, as determined by a radiologist experienced in mammographic interpretation: 1, very likely benign; 10, very likely malignant.

The pixel values were linearly converted before they were stored on the computer so that a high pixel value represented a low optical density.

The location of the biopsied mass was identified by the radiologist, and a region of interest (ROI) containing the biopsied mass was extracted for computerized analysis. The size of the ROI was allowed to vary according to the lesion size. The extracted ROIs contained a non-uniform background, which depended on the overlapping breast structures and the location of the lesion on the mammogram. The non-uniform background is not related to mass malignancy, but may affect the segmentation and feature extraction results used in our computerized analysis. To reduce the background non-uniformity, an automated background correction technique was applied to each ROI as the very first step in our analysis. Details and examples of our background correction technique can be found in the literature (Sahiner *et al* 1996b).

2.2. The rubber-band straightening transform (RBST)

In this study, the classification of malignant and benign masses was based on the textural differences of their mammographic appearance. We have previously designed a rubber-band straightening transform (RBST) which was found to facilitate the extraction of texture features from the region surrounding a mammographic mass. The image transformation performed by the RBST is depicted in figure 2, and a block diagram of different stages of the RBST is given in figure 3. A detailed discussion of the transform can be found in the literature (Sahiner *et al* 1996a, 1997, 1998). For completeness, a brief description is given below.

The RBST transforms a band of pixels surrounding a mass onto the Cartesian plane. The four basic steps in the RBST are mass segmentation, edge enumeration, computation of normals and interpolation. A modified K-means clustering algorithm (Sahiner *et al* 1995) was used for segmentation. The parameters of the segmentation algorithm were chosen so that the segmented region was slightly smaller than the actual size of the mass. After clustering, one to several objects would be segmented in the ROI. If more than one object was segmented, the largest connected object was selected. The selected object

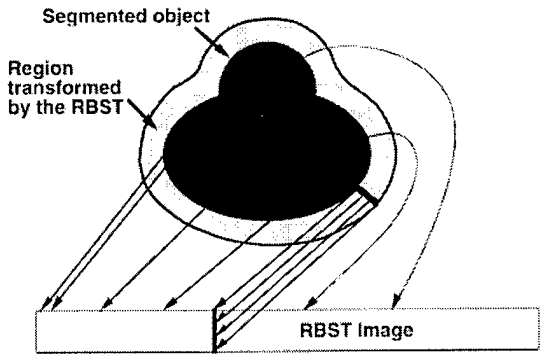


Figure 2. The formation of the RBST image.

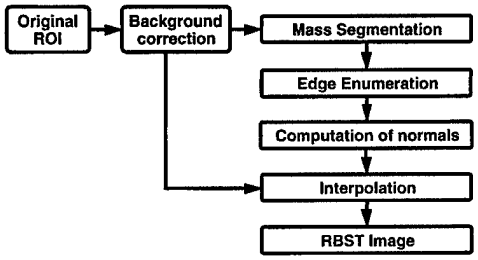


Figure 3. Block diagram of the stages of RBST image computation.

was then filled, grown in a local neighbourhood, and eroded and dilated with morphological operators. The implementation details of these steps have been described elsewhere (Sahiner *et al* 1998). After the outline of the mass was obtained, an edge enumeration algorithm assigned a pixel number to each border pixel of the mass, such that neighbouring pixels were assigned consecutive numbers. The computation of normals depended on the output of the edge enumeration algorithm. The normal $L(i)$ at border pixel i was determined as the normal to the line joining border pixels $i - K$ and $i + K$. The choice of the constant K represents a trade-off between a noisy estimate of the normal direction (small K) and an estimate that misses fine variations in the normal direction (large K). In order to determine the constant K to be used in this study, we selected a small subset of images from our database, and plotted the normal direction obtained by using different values of K superimposed on the segmented image. By performing a visual comparison of the computed normal direction to what was perceived to be the true normal direction, it was empirically found that $K = 12$ resulted in a satisfactory normal estimation. In the interpolation step, the value of the pixel in row j, column i of the RBST image was found as follows. Let $p(i, j)$ denote the location in the original image at a distance j along $L(i)$ from border pixel i. The two closest pixels in the original ROI to location $p(i, j)$ were identified, and the (i, j)th pixel value of the RBST image was defined as the distance-weighted average of these two pixel values.

The width of the band transformed by the RBST was chosen as 40 pixels in this study, which corresponded to 4 mm on the mammogram. An example of the background-corrected ROI, the segmented and morphologically filtered mass shape, and the RBST image are shown in figure 4.

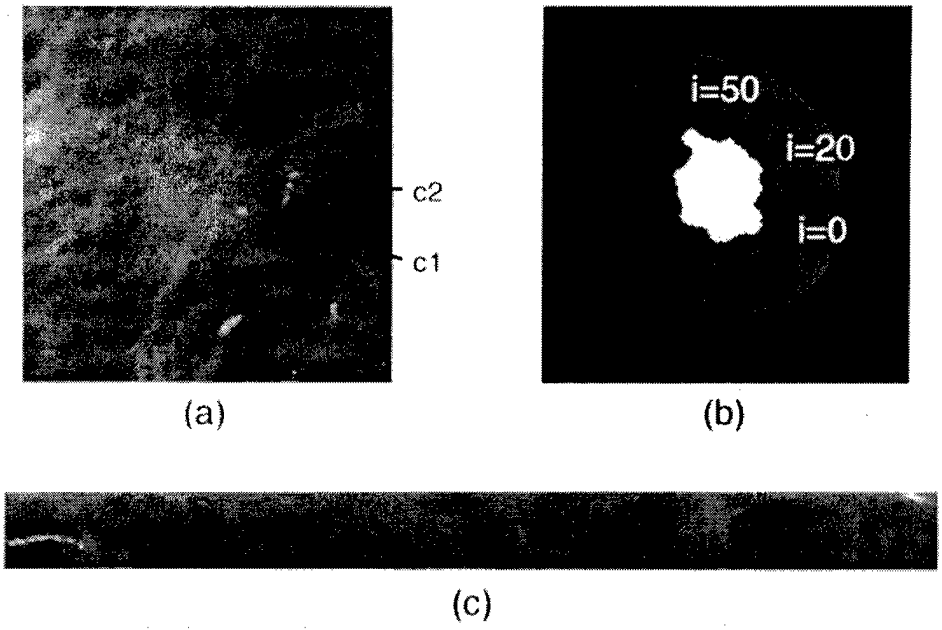


Figure 4. (a) The original mammographic ROI. (b) The segmented and morphologically filtered mass shape (white), and the 40-pixel-wide band around it (grey). For the purpose of illustration, the normals computed at $i = 0$, 20 and 50 are also shown. (c) The RBST image. Notice that due to the position of the first normal location ($i = 0$), the calcifications c1 and c2 on the original ROI appear at the right and the left of the RBST image respectively. The pathological analysis indicated that this was an invasive ductal and intraductal carcinoma.

2.3. Texture features

The texture features used for the classification of the malignant and benign masses were spatial grey-level dependence (SGLD) and run length statistics (RLS) features. These features were extracted from SGLD and RLS matrices, which were constructed from the RBST images as described below.

2.3.1. SGLD features. The (i, j)th element of the SGLD matrix $p_{\theta,d}(i, j)$ represents the probability that grey levels i and j occur at an angle θ and a distance d with respect to each other. The use of SGLD matrices for feature extraction was motivated by the assumption that texture information is contained in the average spatial relationships between the grey-level tones in the image (Haralick *et al* 1973). The features extracted from SGLD matrices of mammographic ROIs have been shown to be useful in classification of mass and normal tissue, and malignant and benign masses or microcalcifications in computer-aided diagnosis (CAD) (Chan *et al* 1995, 1997a, Wei *et al* 1995, Sahiner *et al* 1996b, 1998).

In this study, four different directions ($\theta = 0°$, 45°, 90° and 135°) and ten different pixel pair distances ($d = 1$, 2, 3, 4, 6, 8, 10, 12, 16 and 20) were used for the construction of SGLD matrices from RBST images. The total number of SGLD matrices was therefore 40. Based on our previous studies (Chan *et al* 1995), a bit depth of eight bits was used in the SGLD matrix construction.

A number of SGLD features, which describe the shape of the SGLD matrices, can be extracted from each SGLD matrix. In this study, we extracted eight such features, which

were also used in our previous studies (Chan *et al* 1995, Wei *et al* 1995, Sahiner *et al* 1998). These texture features were correlation, difference entropy, energy, entropy, inertia, inverse difference moment, sum average and sum entropy. This resulted in the computation of 320 SGLD features per RBST image. These features characterize information such as homogeneity, contrast and structural linearity in the images. However, it is difficult to establish a one-to-one correspondence between these qualitative image characteristics and the extracted texture features (Haralick *et al* 1973). The definitions of the SGLD features used in this study can be found in the literature (Haralick *et al* 1973, Chan *et al* 1995, Wei *et al* 1995).

2.3.2. RLS features. The pixels along a given line in an image occasionally contain runs of consecutive pixels that all have the same grey level. A grey-level run is defined as a set of consecutive, collinear pixels in a given direction which have the same grey-level value. A run length is the number of pixels in a grey-level run. The RLS matrix for a given image describes the run length statistics in a given direction for each grey-level value in the image. The (i, j)th element of the RLS matrix $r\theta(i, j)$ represents the number of times that runs of length j in the direction θ consisting of pixels with a grey level i exist in the image (Weszka *et al* 1976).

The RLS matrices in this study were extracted from the vertical and horizontal gradient magnitudes of the RBST images. The vertical and horizontal gradients were obtained by filtering the RBST images with horizontally and vertically oriented Sobel filters (Jain 1989) respectively. Examples of the gradient magnitude images are shown in figure 5. The RLS matrices were obtained from each gradient magnitude image in two directions, $\theta = 0°$ and $\theta = 90°$. Therefore, a total of four RLS matrices were obtained for each RBST image.

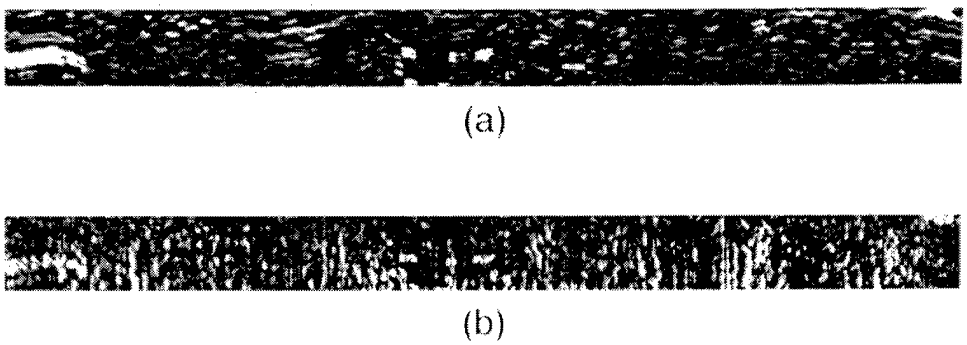


Figure 5. Gradient magnitude images for the RBST image in figure 4: (a) horizontal gradient magnitude image and (b) vertical gradient magnitude image.

Based on our previous study, a bit depth of 5 was used for the computation of RLS matrices (Sahiner *et al* 1998). Five RLS features, namely short runs emphasis, long runs emphasis, grey-level non-uniformity, run length non-uniformity and run percentage were extracted from each RLS matrix. This resulted in the computation of 20 RLS features per RBST image. The definitions of these features can be found in the literature (Galloway 1975). It is possible to describe the general aspects of the relationship between the image characteristics and the RLS feature values. For example, run percentage is low for images with long linear structures, and grey-level non-uniformity is low for images where runs are equally distributed throughout the grey levels (Galloway 1975). However, it is again

difficult to establish a one-to-one correspondence between these texture features and visual image features.

2.4. Fisher's linear discriminant and LDA_{sfs}

For a two-class problem, Fisher's linear discriminant projects the multidimensional feature space onto the real line in such a way that the ratio of between-class sum of squares to within-class sum of squares is maximized after the projection (Duda and Hart 1973). This is the optimal classifier if the features for the two classes have a multivariate Gaussian distribution with equal covariance matrices (Lachenbruch 1975). It has been shown to be a reasonably good classifier even when the feature distributions for the two classes are non-Gaussian (Duda and Hart 1973). Linear discriminant analysis (LDA) is a class of statistical techniques based on Fisher's linear discriminant.

When the training data size is limited, the inclusion of inappropriate features in a classifier may reduce the test accuracy due to overtraining. Therefore, when a large number of features are available for a classification task, it is necessary to select a subset of the most effective features from the feature pool. LDA_{sfs} is a commonly used feature selection method (Lachenbruch 1975). In this study, the performance of a GA-based high-sensitivity feature selection method was compared with that of stepwise feature selection.

Wilks' lambda, which is defined as the ratio of within-group sum of squares to the total sum of squares (Lachenbruch 1975), was used as the selection criterion for the stepwise feature selection method. The stepwise feature selection algorithm starts with no selected features at step 0. At step s of the algorithm, the available features are entered into the selected feature pool one at a time during feature entry, and those already selected are removed one at a time during feature removal. The significance of the change in the Wilks' lambda, as determined by F-statistics, when a new feature is entered into the selected feature pool is compared with a threshold F_{in}. The feature with the highest significance is entered to the selected feature pool only if the significance is higher than F_{in}. Likewise, the significance of the change in the Wilks' lambda when a selected feature is removed from the feature pool is compared with a threshold F_{out}. The feature with the least significance is removed from the selected feature pool only if the significance is lower than F_{out}. This completes step s of the algorithm. The algorithm terminates when no more features can satisfy the criteria for either being added to or removed from the selected feature pool.

2.5. Genetic algorithms for feature selection

Genetic algorithms solve optimization problems by mimicking the natural selection process. A GA follows the evolution of a population of chromosomes which are encoded so that each chromosome corresponds to a possible solution of the optimization problem. The chromosomes consist of genes, which are components of the solution. The goal of a GA is to search for better combinations of the genes, i.e. new chromosomes which are better solutions to the optimization problem. This goal is achieved by evolution. A new generation of chromosomes is produced from the current population by means of parent selection, crossover and mutation. The probability that a chromosome is selected as a parent is related to its ability to solve the optimization problem, i.e. its fitness. Chromosomes which are better solutions to the optimization problem are given a higher chance to reproduce than those which are worse solutions to the problem, similar to the principle of natural selection. The fitness of a chromosome is computed using a fitness function, which is designed on the basis of the optimization criterion for the problem. The probability that a chromosome

is selected as a parent is equal to its normalized fitness, which is defined as the fitness of the chromosome divided by the sum of fitnesses for all chromosomes. The chromosomes of the selected parents are allowed to randomly cross over and mutate, introducing new genes and new chromosomes into the population. This process generates a new population of chromosomes, which tends to evolve towards a better solution.

GAs had been applied to the problem of feature selection (Brill *et al* 1992, Sahiner *et al* 1996c). The most natural way of encoding a chromosome for this problem is as follows (Sahiner *et al* 1996c). Each gene in a chromosome is a bit, which takes a value of either 1 or 0. Each gene location in a chromosome corresponds to a particular feature. If the bit value at a gene location is 1, the corresponding feature is selected for the solution of the classification problem. Otherwise, the corresponding feature is not selected. Each chromosome thus defines a set of selected features. A statistical classifier, such as Fisher's linear classifier or a neural network classifier, is then employed for classification based on the selected feature set. The fitness function reflects the success of the selected feature set for solving the classification problem. The design of the fitness function for a high-sensitivity classifier is described in the next section. The GA training method and the choice of GA parameters are summarized next.

2.5.1. GA training. The GA in this study was trained using a leave-one-case-out paradigm. In this paradigm, all ROIs except those from a particular patient were defined as the training set, and the ROIs from that particular patient were defined as the test set. For each chromosome of the GA, the coefficients of Fisher's linear discriminant function were determined using the features of the training set. The trained discriminant function was then used to classify the test cases using the features of the test cases as the input. In a given generation of the GA, all patients were visited in a round-robin manner, so that test scores were obtained for each ROI in the entire data set. The fitness of the chromosome was computed based on the classification accuracy for the test cases, as described in the next section.

2.5.2. GA parameters. The fundamental parameters of a GA are the number of chromosomes, the chromosome length, the crossover rate, the mutation rate and the stopping criterion. In a GA, the population must contain a large number of chromosomes to provide the variability that offers the opportunity to evolve towards the optimal solution. This requirement and computing speed considerations are trade-offs for selecting the number of chromosomes in a given application. The length of a chromosome is determined by the encoding mechanism which translates the optimization problem into a GA. With the encoding mechanism described earlier in this subsection, the length of each chromosome is equal to the total number of features. The fitness function is the most important component of the GA, and its design is described in the next section. Pairs of chromosomes are probabilistically selected as parents based on their fitness. A selected pair may exchange genes to generate two offspring. The crossover rate determines the probability that parents will exchange genes. After crossover, the binary value of each bit may probabilistically be altered (from 1 to 0, or vice versa), i.e. mutated. The mutation rate determines the probability that genes will undergo mutation. The increase in the fitness of the chromosomes starts to stagnate after a number of generations. The stopping criterion determines when the evolution is terminated. In this study, the GA evolution was terminated after a fixed number of iterations. The appropriateness of this stopping criterion is discussed in section 4. After the termination, the chromosome with the highest fitness value provided the set of selected features.

Table 1 shows the values of each of these parameters, selected based on our previous work. More detailed discussion of these operators and parameters can be found in the literature (Sahiner *et al* 1996c).

Table 1. GA parameters used in this study.

Crossover rate	0.9
Mutation rate	0.0025
Chromosome length	340
Number of chromosomes	200
Stopping criterion	200 iterations

2.6. Design of a high-sensitivity classifier

A widely accepted method for comparing the performance of two classifiers is to consider their ROC curves. The area A_z under the ROC curve is a commonly used index for this comparison. However, for applications where the performance at high sensitivity (or high true-positive fraction) is important, for example breast lesion characterization in CAD, this index may be inadequate. Jiang *et al* (1996) explored this issue, and defined an ROC partial area index that will be denoted as $A_{\mathrm{TPF_0}}$ in this paper.

The partial area index $A_{\mathrm{TPF_0}}$ summarizes the average specificity above a sensitivity of $\mathrm{TPF_0}$ (figure 6), and can be expressed as (Jiang *et al* 1996)

$$A_{\mathrm{TPF_0}} = 1 - \frac{1}{1-\mathrm{TPF_0}} \int_{\mathrm{TPF_0}}^{1} \mathrm{FPF(TPF)}\, \mathrm{d(TPF)} \tag{1}$$

which is the ratio of the partial area under the actual ROC curve to the partial area of the perfect ROC curve. The maximum value for $A_{\mathrm{TPF_0}}$ is thus 1. The $A_{\mathrm{TPF_0}}$ value for a classifier that operates purely on random guessing is $(1-\mathrm{TPF_0})/2$, which is the area under the chance diagonal normalized to $1-\mathrm{TPF_0}$.

When the conventional binormal model is employed for the computation of the ROC curve, the curve is completely defined by two parameters, a and b, which are determined from the rating data using maximum likelihood estimation. The constant b represents the estimated standard deviation of the actually negative cases, normalized by the estimated standard deviation of the actually positive cases, and the constant a represents the estimated difference between the means of actually positive and negative cases, normalized again by the estimated standard deviation of the actually positive cases. Using the binormality assumption, the partial area index $A_{\mathrm{TPF_0}}$ can be expressed as (McClish 1989, Jiang *et al* 1996)

$$A_{\mathrm{TPF_0}} = 1 - \frac{1}{1-\mathrm{TPF_0}} \int_{c_0}^{\infty} \Phi\left(\frac{u-a}{b}\right) \phi(u)\, \mathrm{d}u \tag{2}$$

where

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$$

and

$$\Phi(u) = \int_{-\infty}^{u} \phi(x)\, \mathrm{d}x.$$

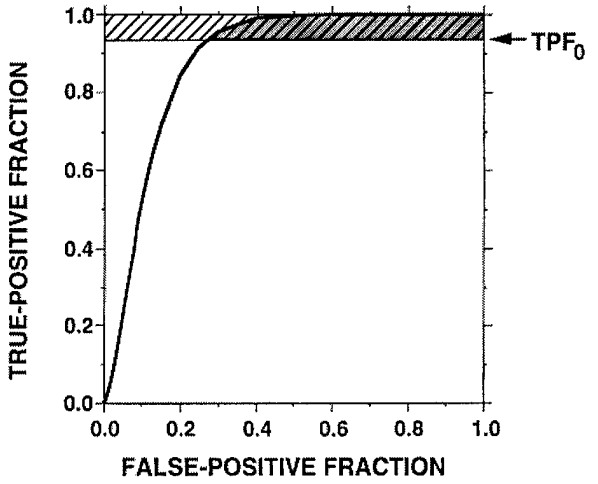


Figure 6. The partial area index A_{TPF_0} is defined as the ratio of the partial area under the ROC curve above a given sensitivity (grey area) to the partial area of the perfect ROC curve (hatched region) above the same sensitivity.

Our goal in this study was to train a GA to select features which would yield high specificity in the high-sensitivity region of the ROC curve. Therefore, the fitness of a chromosome was defined as a monotonic function of A_{TPF_0}, such that the maximization of A_{TPF_0} would maximize the fitness function

$$\text{fitness} = \left(\frac{A_{TPF_0} - A_{min}}{A_{max} - A_{min}} \right)^n \tag{3}$$

where A_{max} and A_{min} were the maximum and minimum values of A_{TPF_0} among all chromosomes in a generation, and n was a power parameter whose effect on GA feature selection was investigated, as discussed in section 3. From equation (3), it is seen that as the power parameter becomes larger the difference in the fitness, and thus the probability of being chosen as parents, between the chromosomes are more amplified. The choice of n is a tradeoff between the goal of promoting chromosomes with high fitness values and the need to retain segments of good genes in other chromosomes.

For a given chromosome, the parameters a and b that are required for the computation of A_{TPF_0} were determined from the distribution of test scores using the LABROC program of Metz *et al* (1998). The partial area index A_{TPF_0} was then computed by numerically integrating equation (2). The classifiers thus designed will be referred to as GA-based high-sensitivity classifiers in the following discussions.

In this study, the significance of the difference in A_{TPF_0} of different classifiers was determined using a recently developed statistical test (Jiang *et al* 1996). The test is analogous to statistical tests involving the area A_z under the entire ROC curve, and is implemented using the covariance estimates of a and b values for the two curves.

3. Results

To demonstrate the training of high-sensitivity classifiers using GA, we chose two levels of sensitivity thresholds, $TPF_0 = 0.50$ and $TPF_0 = 0.95$ in equation (1). The classification results of these classifiers were compared with those of LDA_{sfs}. GA-based feature selection

Table 2. The number of features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) for various values of F_{in} and F_{out} in the stepwise feature selection method.

F_{in}	F_{out}	Number of selected features	A_z	$A_{0.50}$	$A_{0.95}$
3.8	2.7	9	0.84	0.71	0.22
2.6	2.4	13	0.85	0.72	0.27
2.2	2.0	14	0.86	0.73	0.25
1.8	1.6	26	0.89	0.80	0.38
1.4	1.2	41	0.92	0.83	0.47
1.0	1.0	49	0.92	0.83	0.46

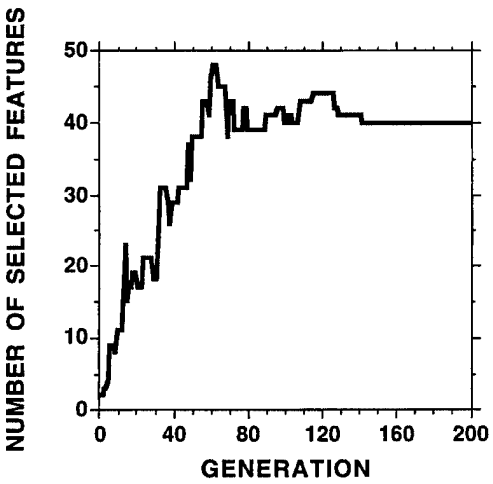


Figure 7. The evolution of the number of selected features for a GA training session ($n = 4$, $TPF_0 = 0.95$).

was also performed with no emphasis on high sensitivity ($TPF_0 = 0$). The classifier designed with the features thus selected will be referred to as an ordinary GA-based classifier. Its performance was compared with those of the GA-based high-sensitivity classifiers and LDA_{sfs}.

In LDA_{sfs}, the optimal values of the F_{in} and F_{out} thresholds are not known *a priori*. We therefore varied these thresholds to obtain the feature subset with the best test performance. Table 2 shows the number of selected features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) as these F thresholds are varied. By comparing the A_z values and the performance at the high-sensitivity portion of the ROC curve, the combination $F_{in} = 1.4$, $F_{out} = 1.2$ was found to provide the best feature subset.

High-sensitivity classifiers with $TPF_0 = 0.50$ and $TPF_0 = 0.95$ were trained with three different values of the power parameter, n ($n = 1$, 2 and 4). Figure 7 shows the evolution of the number of selected features, and figure 8 shows the total area under the ROC curve (A_z) and the partial area above the true positive fraction of 0.95 ($A_{0.95}$) for a typical GA training ($n = 4$, $TPF_0 = 0.95$).

The ROC curve of the best LDA_{sfs} classifier and those of GA-based classifiers ($TPF_0 = 0.50$ and $TPF_0 = 0.95$) with $n = 1$, 2 and 4 are compared in figures 9–11

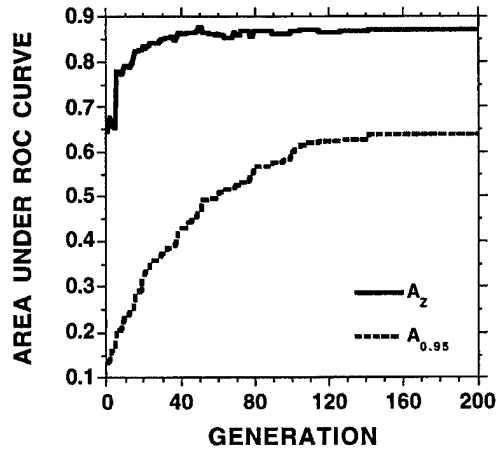


Figure 8. The evolution of the area A_z and the partial area $A_{0.95}$ under the ROC curve for the GA training session of figure 7 ($n = 4$, $TPF_0 = 0.95$).

respectively. It is observed from figures 10 and 11 that for $n = 2$ or 4, the designed high-sensitivity classifiers seem to be superior to the best LDA_{sfs} classifier for large values of true positives. When $n = 1$, the ROC curves of the GA-based high-sensitivity classifiers are still higher than that of the LDA_{sfs} classifier when TPF is very close to 1; however, the difference between the curves is small. To quantify the improvement obtained by the GA-based high-sensitivity classifier, we performed statistical significance tests (Jiang *et al* 1996) on the partial area above a true-positive threshold of 0.95 ($A_{0.95}$) as described in the previous section. With $n = 4$, the difference between the partial areas of the GA-based high-sensitivity classifiers and LDA_{sfs} above a true-positive threshold of 0.95 was statistically significant with two-tailed p-levels of 0.006 and 0.02 for the classifiers trained with $TPF_0 = 0.95$ and $TPF_0 = 0.5$ respectively. For $n = 2$, the corresponding p-levels were 0.01 and 0.07 respectively. For $n = 1$, the difference did not achieve statistical significance ($p = 0.14$ for $TPF_0 = 0.95$ and $p = 0.49$ for $TPF_0 = 0.5$). The difference of the partial area index over a true-positive threshold of 0.5 ($A_{0.50}$) did not achieve statistical significance when the high-sensitivity classifiers trained with $TPF_0 = 0.5$ were compared with LDA_{sfs} for any of the power parameters studied ($n = 1$, 2 and 4).

The performance of the high-sensitivity classifiers and the ordinary GA-based classifiers ($TPF_0 = 0$) are also compared in figures 9–11. It is observed that the difference between the high-sensitivity and the ordinary GA-based classifiers is less than the difference between the high-sensitivity classifiers and the LDA_{sfs}. With a two-tailed significance test, it was found that the difference between the partial areas of the high-sensitivity and the ordinary GA-based classifiers above a true-positive threshold of 0.95 ($A_{0.95}$) did not achieve statistical significance for any of the power parameter values studied ($n = 1$, 2 and 4) with p-levels ranging between 0.06 and 0.5. Similarly, the difference between the ordinary GA-based classifiers and LDA_{sfs} did not achieve statistical significance for any of the power parameter values studied. Table 3 summarizes the A_z, $A_{0.50}$ and $A_{0.95}$ values, as well as the number of features selected by each classifier.

Figures 12 and 13 show the distributions of the classifier outputs for the high-sensitivity classifier ($n = 4$, $TPF_0 = 0.95$) and the LDA_{sfs} respectively. Using the LDA_{sfs}, the distribution of the malignant masses has a relatively long tail that overlaps with the distribution of the benign masses. With the high-sensitivity classifier, this tail seems to

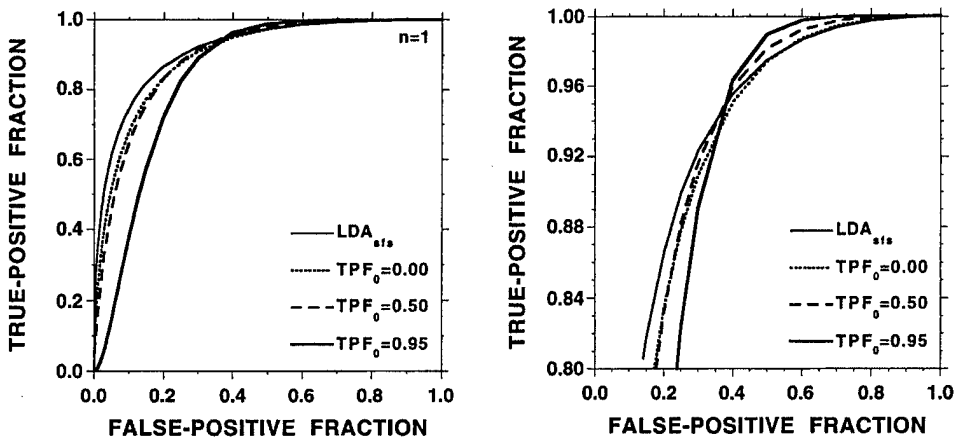


Figure 9. The ROC curves of the LDA_{sfs}, the ordinary GA-based classifier ($TPF_0 = 0$), and the GA-based high-sensitivity classifiers trained with $TPF_0 = 0.50$ and $TPF_0 = 0.95$ using power parameter $n = 1$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.

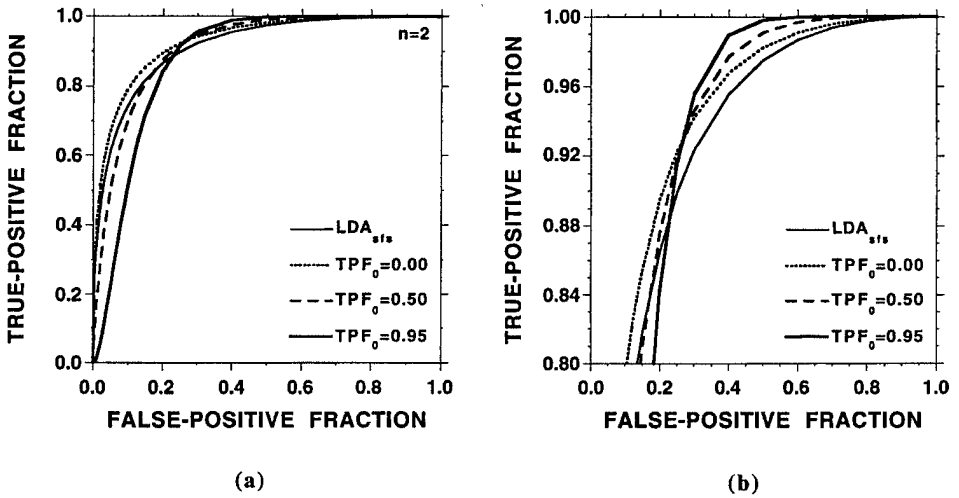


Figure 10. The ROC curves of the LDA_{sfs}, the ordinary GA-based classifier ($TPF_0 = 0$), and the GA-based high-sensitivity classifiers trained with $TPF_0 = 0.50$ and $TPF_0 = 0.95$ using power parameter $n = 2$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.

be shortened, so that more benign masses may be correctly diagnosed without missing malignancies. At 100% sensitivity, the specificity with the appropriate choice of the decision threshold was 61% and 34% for the high-sensitivity classifier and the LDA_{sfs} respectively.

4. Discussion

Figures 10 and 11 demonstrate that when the feature selection is performed with a properly designed fitness function in the GA, the designed classifier can be more effective than

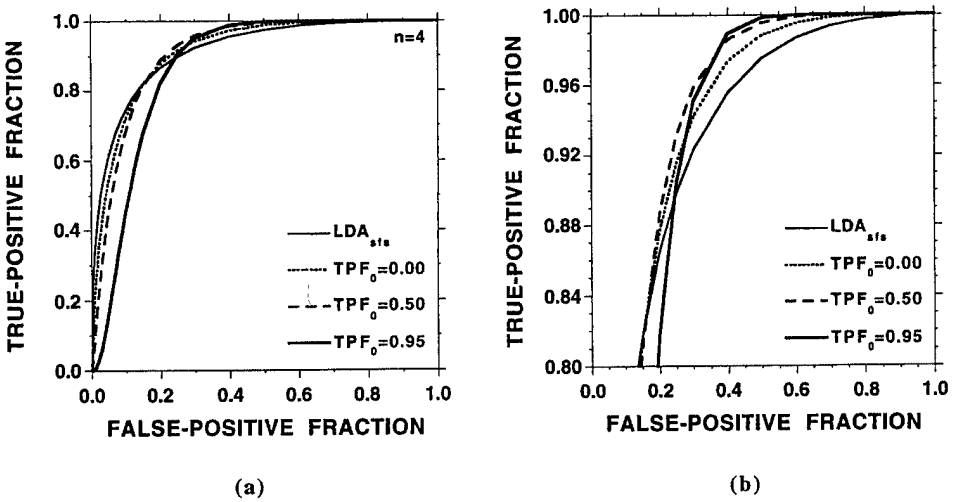


Figure 11. The ROC curves of the LDA_{sfs}, the ordinary GA-based classifier ($TPF_0 = 0$), and the GA-based high-sensitivity classifiers trained with $TPF_0 = 0.50$ and $TPF_0 = 0.95$ using power parameter $n = 4$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.

Table 3. The number of features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) for the GA parameters studied. For comparison purposes, the results with linear discriminant analysis are also included as the last row.

Power Parameter, n	TPF_0 value for GA training	Number of selected features	A_z	$A_{0.50}$	$A_{0.95}$
1	0	62	0.90 ± 0.02	0.81 ± 0.03	0.47 ± 0.07
1	0.5	61	0.89 ± 0.02	0.81 ± 0.03	0.51 ± 0.07
1	0.95	58	0.84 ± 0.02	0.76 ± 0.03	0.55 ± 0.05
2	0	60	0.93 ± 0.02	0.86 ± 0.03	0.51 ± 0.08
2	0.5	48	0.91 ± 0.02	0.85 ± 0.03	0.58 ± 0.07
2	0.95	50	0.88 ± 0.02	0.82 ± 0.03	0.63 ± 0.05
4	0	40	0.92 ± 0.02	0.85 ± 0.03	0.56 ± 0.07
4	0.5	39	0.91 ± 0.02	0.85 ± 0.03	0.62 ± 0.06
4	0.95	40	0.87 ± 0.02	0.81 ± 0.03	0.64 ± 0.05
Linear discriminant analysis		41	0.92 ± 0.02	0.83 ± 0.03	0.47 ± 0.07

LDA_{sfs} in the high-sensitivity region of the ROC curve. From table 3 it is observed that although the A_z value for the properly trained high-sensitivity classifier (e.g. $TPF_0 = 0.5$ or 0.95 and $n = 2$ or 4) may be less than that of the LDA_{sfs}, the partial area index $A_{0.95}$ is larger. The statistical analysis in this study showed that the difference between the properly designed high-sensitivity classifiers and the LDA_{sfs} at the high-sensitivity region of the ROC curve can be significant.

Comparing figure 9 with figures 10 and 11, it is observed that the selection of the power parameter n in GA training may be important. The classifiers designed with $n = 1$ did not exhibit a major advantage over the LDA_{sfs}, as also seen from table 3 and the statistical significance tests. From equation (3), it is seen that as the power parameter becomes larger, the difference in the fitness, and thus the probability of being chosen as

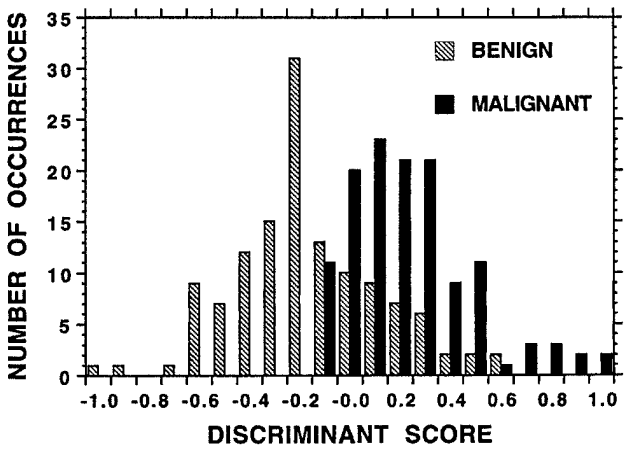


Figure 12. The distribution of the classifier output for the high-sensitivity classifier with $n = 4$, $TPF_0 = 0.95$. By setting an appropriate threshold on these classifier scores, 61% of masses could correctly be classified as benign without missing any malignancies in this study.

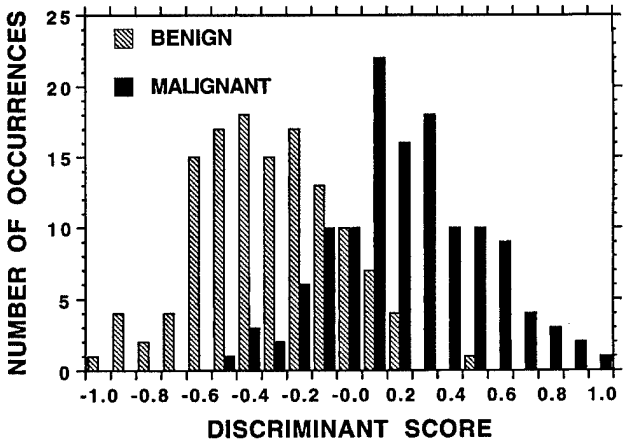


Figure 13. The distribution of the classifier output for LDA_{sfs}. By setting an appropriate threshold on these classifier scores, 34% of masses could be correctly classified as benign without missing any malignancies in this study.

parents, between the chromosomes are more amplified. Therefore, a larger value of n favours the reproduction of better chromosomes in a generation. Although it is desirable to favour the better chromosomes in any GA algorithm, too much emphasis on better chromosomes might suppress the chance of retaining segments of good genes in other chromosomes in the gene pool. This is best seen by letting n tend to infinity, and observing that only the best single chromosome will reproduce in this case, which reduces the GA to a random search algorithm. In our application, from table 3, it is observed that, for all three sensitivity thresholds ($TPF_0 = 0.95$, 0.50 and 0), the classifier trained with $n = 1$ has lower performance indices ($A_{0.95}$, $A_{0.50}$ and A_z) than its counterpart trained with $n = 2$ or $n = 4$. Although none of these differences reached statistical significance, the consistently poorer performance of the classifiers trained with $n = 1$ indicates that $n = 1$ may not be a good choice for GA training.

From figures 7 and 8 it is observed that the best fitness and the number of chromosomes did not change between iterations 140 and 200 for the high-sensitivity classifier with $n = 4$ and $TPF_0 = 0.95$. A similar trend was observed with the other values of n and TPF_0 investigated in this study. Therefore, 200 generations seems to be sufficient for the GA to complete its evolution in this application. In figure 8, the best A_z value was attained around the fiftieth generation, and the A_z value did not change considerably afterwards. However, the $A_{0.95}$ value increased until around 140 generations. This meant that the classification accuracy at high sensitivity continued to increase although the A_z value did not change, i.e. the shape of the ROC curve changed so that the specificity at the high-sensitivity region of the ROC curve increased, while the specificity at the low-sensitivity region of the ROC curve decreased.

Figures 9–11 and the statistical significance tests in section 3 show that although the GA-based high-sensitivity classifiers perform better than the ordinary GA-based classifiers at high sensitivity, the difference between the two classifiers is not statistically significant. Comparison of the LDA_{sfs} and the ordinary GA-based classifiers revealed that neither the difference between the A_z values, nor the difference between the $A_{0.95}$ values were statistically significant ($p > 0.3$). However, the difference between the $A_{0.95}$ values of the LDA_{sfs} and the GA-based high-sensitivity classifiers trained with power parameter $n = 2$ and $n = 4$ was statistically significant (two-tailed p-level <0.05), as described in section 3. Thus, it was necessary to use a high-sensitivity classifier in order to obtain statistically significant improvement over the LDA_{sfs}.

The GA-based high-sensitivity classifiers ($TPF_0 = 0.95$ and $TPF_0 = 0.5$) and the ordinary GA-based classifier ($TPF_0 = 0$) were designed to maximize the partial ROC areas above the chosen true-positive fraction thresholds. From table 3, it is observed that this goal is achieved for the GA-based classifiers with TPF_0 values of 0 and 0.95. For each n, the GA-based classifier with $TPF_0 = 0$ (ordinary GA-based classifier) yielded the highest A_z value, and the GA-based classifier with $TPF_0 = 0.95$ yielded the highest $A_{0.95}$ value among the classifiers. For the classifier with $TPF_0 = 0.5$, the $A_{0.50}$ value was larger than or equal to that of the other GA-based classifiers for $n = 1$ and $n = 4$. However, for $n = 2$, the ordinary GA-based classifier ($TPF_0 = 0$) had the highest $A_{0.50}$ value, although the difference was not statistically significant ($p > 0.3$). This result is not inconsistent with the GA principles or operation. Since the GA training is based on stochastic search, the GA tends to evolve towards the optimal solution, as evidenced by the comparison of the GA-based classifiers in table 3. However, the optimality of the solution is not guaranteed, and one may encounter situations that the design goal was not totally achieved, as evidenced by the fact that the ordinary GA-based classifier had the highest $A_{0.50}$ value for $n = 2$.

Given the probabilistic nature of GA-based feature selection, it is difficult to predict the conditions under which the GA may select a feature set that provides a better high-sensitivity classifier than LDA_{sfs}. Both our GA-based method and the stepwise feature selection algorithm were designed primarily to select features for classifying classes that have multivariate Gaussian distributions and equal covariance matrices. When these assumptions are not satisfied, the accuracy of feature selection will deteriorate to a different degree for both methods. One possible explanation for the relative success of the GA-based feature selection might be that our data violate the assumptions of multivariate normality and the equality of covariance matrices, and that the GA-based method is less sensitive to these violations.

In this study, our focus was to develop a methodology for the design of high-sensitivity classifiers for applications in CAD. For the specific application of discriminating malignant and benign breast lesions, our data set was limited and the features selected by the GA

may not be the optimal set of features for the general population. The same is true for the LDA_{sfs}. Considering that the data set contained only 255 masses, the number of features selected both by the GA and the LDA_{sfs} was large. As a result, if a classifier trained in this study is applied without modification to the population at large, the classification accuracy is likely to be poorer than that obtained in this paper. However, the methodology developed in this study is general. When a sufficiently large data set becomes available, the GA-based high-sensitivity feature selection algorithm can be reapplied, and a more robust feature set can be determined. The number of training cases required for generalizable classifier design and feature selection has been the subject of recent studies (Raudys and Jain 1991, Wagner *et al* 1997, Chan *et al* 1997b), and is currently under investigation.

An important consideration concerning the use of GAs for optimization is the speed of computation. Depending on the number of final features selected, the GA-based feature selection implemented in this study (340 features, 200 chromosomes, 200 generations and leave-one-case-out GA training) took between 24 and 60 h on an AlphaStation 500 (400 Mhz Alpha chip), whereas the stepwise feature selection performed on a PC compatible computer with a 90 MHz Pentium processor took less than 10 min. Therefore, GA-based feature selection implemented in this study may not be practical for studies where the feature selection has to be performed many times. The high-sensitivity classifier design method developed in this study may be more appropriate if the speed of computation is of secondary importance to the classification accuracy of the designed classifier. For example, the GA-based high-sensitivity classifier can be trained only once when a final set of features is desired for a large data set as discussed above.

5. Conclusion

We have developed a GA-based method to design a high-sensitivity classifier for CAD applications. The usefulness of the method was demonstrated by the problem of classifying masses on digitized mammograms. Texture features extracted from RBST images were used to distinguish malignant and benign masses. The accuracy of the high-sensitivity classifier was shown to be significantly higher than that of LDA_{sfs} above a true-positive fraction of 0.95. By using an appropriate decision threshold on the high-sensitivity classifier scores, 61% of the benign masses could correctly be identified without missing any malignant masses. The GA may therefore be a useful tool in the design of high-sensitivity classifiers for different classification problems in CAD or other applications.

Acknowledgments

This work is supported by a USPHS grant CA 48129, a Career Development Award (BS) from the USAMRMC (DAMD 17-96-1-6012), and a USAMRMC grant DAMD 17-96-1-6254. No official endorsement of any equipment or product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E Metz, PhD, for providing the LABROC and CLABROC programs.

References

Brill F, Brown D and Martin W 1992 Fast genetic selection of features for neural network classifiers *IEEE Trans. Neural Networks* **3** 324–8

Brzakovic D, Luo X M and Brzakovic P 1990 An approach to automated detection of tumors in mammograms *IEEE Trans. Med. Imaging* **9** 233–41

Chan H-P, Sahiner B, Petrick N, Helvie M A, Lam K L, Adler D D and Goodsitt M M 1997a Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network *Phys. Med. Biol.* **42** 549–67

Chan H-P, Sahiner B, Wagner R F, Petrick N and Mossoba J 1997b Effects of sample size on classifier design: quadratic and neural network classifiers *Proc. SPIE* **3034** 1102–13

Chan H-P, Wei D, Helvie M A, Sahiner B, Adler D D, Goodsitt M M and Petrick N 1995 Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space *Phys. Med. Biol.* **40** 857–76

Duda R O and Hart P E 1973 *Pattern Classification and Scene Analysis* (New York: Wiley)

Galloway M M 1975 Texture classification using grey level run lengths *Comput. Graphics Image Process.* **4** 172–9

Hall F M, Storella J M, Silverstone D Z and Wyshak G 1988 Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography *Radiology* **167** 353–8

Haralick R M, Shanmugam K and Dinstein I 1973 Texture features for image classification *IEEE Trans. Systems Man Cybernetics* **3** 610–21

Hermann G, Janus C, Schwartz I S, Krivisky B, Bier S and Rabinowitz J G 1987 Nonpalpable breast lesions: accuracy of prebiopsy mammographic diagnosis *Radiology* **165** 323–6

Huo Z, Giger M L, Vyborny C J, Bick U, Lu P, Wolverton D E and Schmidt R A 1995 Analysis of spiculation in the computerized classification of mammographic masses *Med. Phys.* **22** 1569–79

Jacobson H G and Edeiken J 1990 Biopsy of occult breast lesions: analysis of 1261 abnormalities *J. Am. Med. Assoc.* **263** 2341–3

Jain A K 1989 *Fundamentals of Digital Image Processing* (New Jersey: Prentice-Hall)

Jiang Y, Metz C E and Nishikawa R M 1996 A receiver operating characteristic partial area index for highly sensitive diagnostic tests *Radiology* **201** 745–50

Kilday J, Palmieri F and Fox M D 1993 Classifying mammographic lesions using computerized image analysis *IEEE Trans. Med. Imaging* **12** 664–9

Lachenbruch P A 1975 *Discriminant Analysis* (New York: Hafner)

McClish D K 1989 Analyzing a portion of the ROC curve *Med. Decision Making* **9** 190–5

Metz C E, Herman B A and Shen J H 1998 Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data *Stat. Med.* **17** 1033–53

Pohlman S, Powell K A, Obuchowski N A, Chilcote W A and Broniatowski S G 1996 Quantitative classification of breast tumors in digitized mammograms *Med. Phys.* **23** 1337–45

Rangayyan R M, El-Faramawy N, Desautels J E L and Alim O A 1996 Discrimination between benign and malignant breast tumors using a region-based measure of edge profile acutance *Digital Mammography '96* ed K Doi, M L Giger, R M Nishikawa and R A Schmidt (Amsterdam: Elsevier) pp 213–18

Raudys S J and Jain A K 1991 Small sample size effects in statistical pattern recognition: recommendations for practitioners *IEEE Trans. Pattern Anal. Machine Intell.* **13** 252–64

Sahiner B, Chan H-P, Petrick N, Goodsitt M M and Helvie M A 1997 Characterization of masses on mammograms: significance of the use of the rubber-band straightening transform *Proc. SPIE* **3034** 491–500

Sahiner B, Chan H-P, Petrick N, Helvie M A and Goodsitt M M 1998 Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis *Med. Phys.* **25** 516–26

Sahiner B, Chan H-P, Petrick N, Helvie M A, Goodsitt M M and Adler D D 1996a Classification of masses on mammograms using a rubber-band straightening transform and feature analysis *Proc. SPIE* **2710** 44–50

Sahiner B, Chan H-P, Petrick N, Wei D, Helvie M A, Adler D D and Goodsitt M M 1995 Classification of mass and normal breast tissue: an artificial neural network with morphological features *Proc. World Congress on Neural Networks* vol 2 (New Jersey: INNS Press) pp 876–9

——1996b Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images *IEEE Trans. Med. Imaging* **15** 598–610

——1996c Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue *Med. Phys.* **23** 1671–84

Wagner R F, Chan H-P, Mossoba J, Sahiner B and Petrick N 1997 Finite-sample effects and resampling plans: application to linear classifiers in computer-aided diagnosis *Proc. SPIE* **3034** 467–77

Wei D, Chan H-P, Helvie M A, Sahiner B, Petrick N, Adler D D and Goodsitt M M 1995 Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis *Med. Phys.* **22** 1501–13

Weszka J S, Dyer C R and Rosenfeld A 1976 A comparative study of texture measures for terrain classification *IEEE Trans. Syst. Man Cybernetics* **6** 269–85

Improvement of Radiologists' Characterization of Mammographic Masses by Using Computer-aided Diagnosis: An ROC Study[1]

Heang-Ping Chan, PhD
Berkman Sahiner, PhD
Mark A. Helvie, MD
Nicholas Petrick, PhD
Marilyn A. Roubidoux, MD
Todd E. Wilson, MD
Dorit D. Adler, MD
Chintana Paramagul, MD
Joel S. Newman, MD
Sethumadavan Sanjay-Gopal, PhD

Index terms:
Breast neoplasms, 00.31, 00.32
Breast neoplasms, radiography, 00.111, 00.119
Breast radiography, 00.111, 00.119
Computers, diagnostic aid
Receiver operating characteristic curve (ROC)

Radiology 1999; 212:817–827

Abbreviations:
CAD = computer-aided diagnosis
PPV = positive predictive value
ROC = receiver operating characteristic

[1] From the Department of Radiology, University of Michigan Hospital, UH B1F510, 1500 E Medical Center Dr, Ann Arbor, MI 48109-0030. From the 1997 RSNA scientific assembly. Received August 10, 1998; revision requested September 8; revision received November 30; accepted January 21, 1999. Supported in part by United States Public Health Service grant CA 48129 and by U.S. Army Medical Research and Materiel Command grant DAMD 17-96-1-6254. B.S. supported by Career Development award DAMD 17-96-1-6012 from the U.S. Army Medical Research and Materiel Command. N.P. supported by a grant from the Whitaker Foundation. **Address reprint requests to** H.P.C. (e-mail: *chanhp@umich.edu*).

The content of this article does not necessarily reflect the position of the funding agencies, and no official endorsement of any equipment or product of any companies mentioned in this article should be inferred.

Author contributions:
Guarantor of integrity of entire study, H.P.C.; study concepts and design, H.P.C., M.A.H., B.S., N.P.; literature research, H.P.C., M.A.H.; experimental studies, M.A.H., M.A.R., T.E.W., D.D.A., C.P., J.S.N.; data acquisition, all authors; data analysis, H.P.C., B.S., N.P.; statistical analysis, H.P.C.; manuscript preparation, editing, and review, H.P.C., B.S., M.A.H., N.P., M.A.R., T.E.W., D.D.A., C.P., J.S.N.

PURPOSE: To evaluate the effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses seen on mammograms.

MATERIALS AND METHODS: The authors previously developed an automated computer program for estimation of the relative malignancy rating of masses. In the present study, the authors conducted observer performance experiments with receiver operating characteristic (ROC) methodology to evaluate the effects of computer estimates on radiologists' confidence ratings. Six radiologists assessed biopsy-proved masses with and without CAD. Two experiments, one with a single view and the other with two views, were conducted. The classification accuracy was quantified by using the area under the ROC curve, A_z.

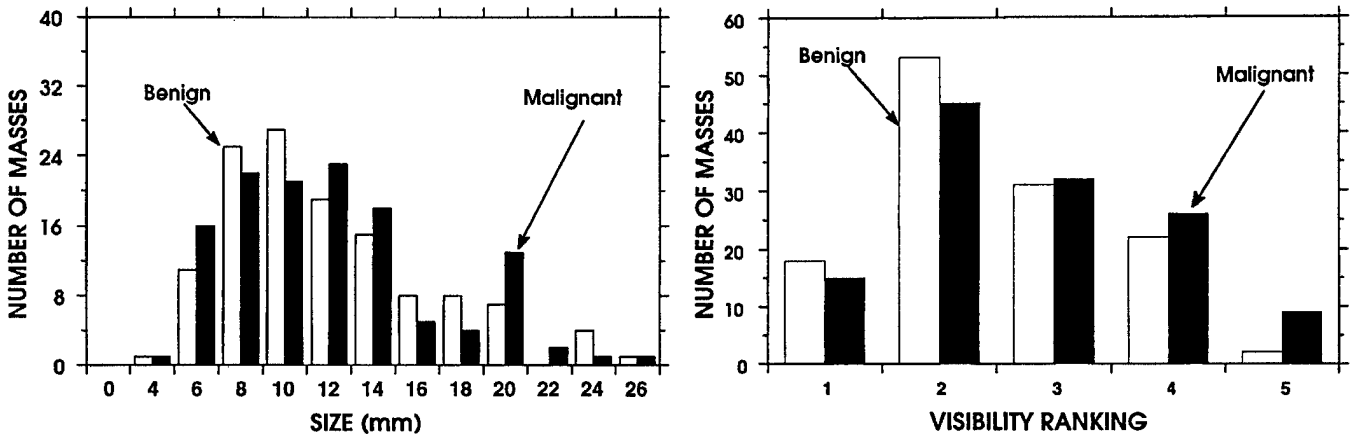
RESULTS: For the reading of 238 images, the A_z value for the computer classifier was 0.92. The radiologists' A_z values ranged from 0.79 to 0.92 without CAD and improved to 0.87–0.96 with CAD. For the reading of a subset of 76 paired views, the radiologists' A_z values ranged from 0.88 to 0.95 without CAD and improved to 0.93–0.97 with CAD. Improvements in the reading of the two sets of images were statistically significant (P = .022 and .007, respectively). An improved positive predictive value as a function of the false-negative fraction was predicted from the improved ROC curves.

CONCLUSION: CAD may be useful for assisting radiologists in classification of masses and thereby potentially help reduce unnecessary biopsies.

Breast cancer is the most prevalent non-skin cancer in women; 178,700 new cases are estimated to have occurred in 1998 (1). The mortality of breast cancer is the second highest among all cancer deaths in women (1). At present, there is no effective method to prevent breast cancer. The best approach to reducing the breast cancer mortality rate is early detection and treatment. Because the mammographic features of early-stage breast cancers are not very specific, the need for high detection sensitivity leads to biopsy of many low-suspicion lesions. The positive predictive values (PPVs) of mammographic signs are, therefore, often below 30% (2,3).

Computer-aided diagnosis (CAD) is considered to be one of the approaches that may improve the efficacy of mammography (4). With CAD, a computerized detection algorithm alerts a radiologist to the location of the suspicious lesions, and/or a trained computer classifier provides the radiologist with an estimate of the likelihood of malignancy of a lesion. The radiologist takes into consideration the information provided by the computer before making a decision. This "second opinion" may improve the diagnostic accuracy because it serves as a form of double reading (5). Furthermore, a computer evaluation is often more consistent and reproducible than a human decision maker (6).

Considerable research has been devoted to the development of computerized schemes for the detection and classification of mammographic abnormalities. These efforts have advanced the CAD technology such that clinical application appears to be possible in the



a. b.

Figure 1. Histograms illustrate the distributions of **(a)** size (ie, length of the long axis) and **(b)** visibility ranking (1 = obvious, 5 = subtle) of the 253 masses included in the data set. Because classification accuracy depends on the case mix, these distributions provided some information on the masses in the data set.

near future. It is, therefore, necessary to evaluate the effects of CAD on radiologists' detection and diagnosis of mammographic lesions. In a previous receiver operating characteristic (ROC) study, we demonstrated that CAD could improve radiologists' accuracy in the detection of subtle microcalcifications on mammograms (7). Kegelmeyer et al (8) also reported an improvement in radiologists' sensitivity for the detection of spiculated masses with use of a computer aid. For the classification of mammographic lesions, it has been shown that a computer classifier that estimated the likelihood of malignancy on the basis of mammographic features extracted by radiologists could improve radiologists' accuracy in distinguishing malignant from benign lesions (9–11).

We previously conducted ROC studies to compare the performance of radiologists with that of the computer (12) and to compare radiologists' ability to classify masses with and without CAD (13). Jiang et al (14) also performed an ROC study of the effect of CAD on radiologists' performance in classifying microcalcifications. The results of all of these observer performance studies indicate the potential to improve mammographic interpretation with a computer aid.

We have developed an automated method to analyze masses seen on mammograms (15–17). A mass is segmented from its surrounding breast tissue, and an image transformation technique is used to transform the mass margin from the polar coordinate system to the Cartesian coordinate system. A linear discriminant classifier then extracts the useful texture features from the transformed image and merges them into a relative malignancy rating. Our approach is different from others that use a trained classifier to merge radiologist-extracted image features or feature codes by using the American College of Radiology Breast Imaging Reporting and Database System lexicon (9–11). Our fully automated method has the advantage that, unlike a human reader, it does not have variability in feature recognition and coding. In addition, the computer may be able to extract some information, such as texture features, that may not be readily perceived by human eyes. We conducted an ROC study to evaluate whether this computer aid can improve radiologists' performance in the classification of mammographic masses (13). The results of our observer performance study are described in this article.

Other investigators also have reported on automated algorithms for the classification of mammographic masses (18–21). The methods used in these algorithms varied, and their accuracy in classification cannot be compared directly because of the differences in the data sets. However, the effects of CAD on radiologists' performance are not expected to depend strongly on the specific algorithm if different computer aids of comparable accuracy are used. Therefore, the applications of the findings of this study should not be limited to our computerized classification aid.

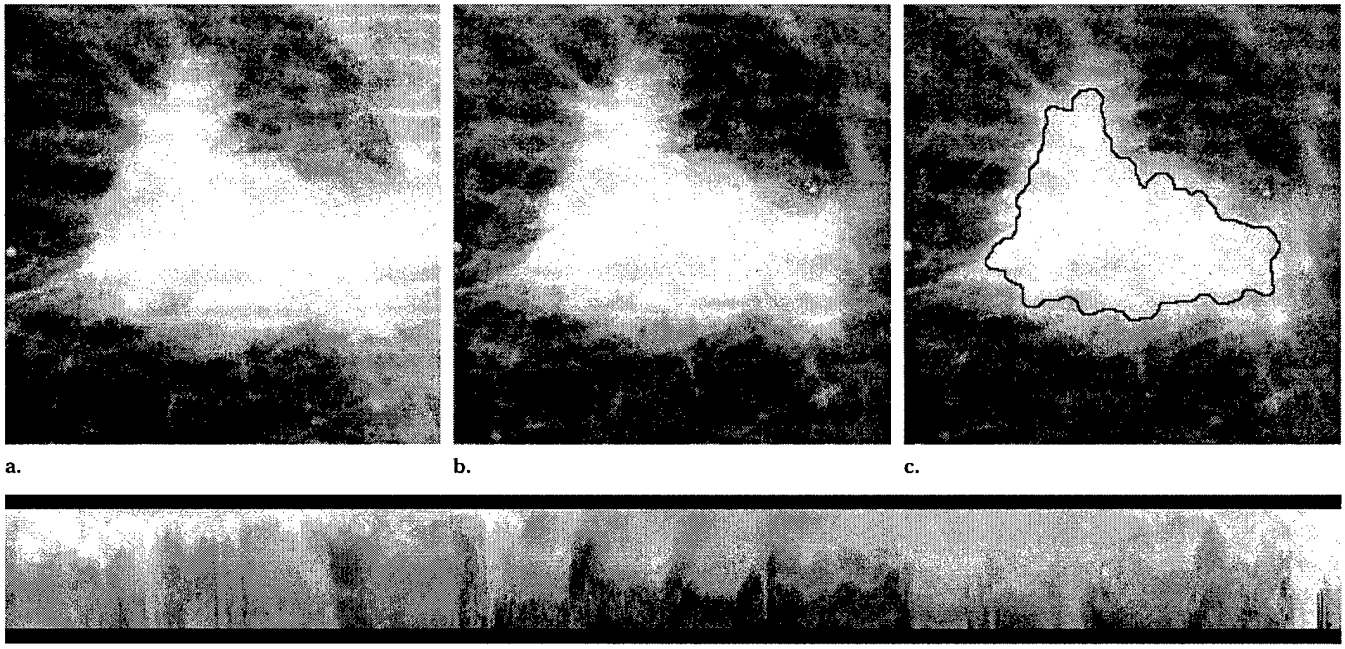
MATERIALS AND METHODS

Data Set

The data set for this study consisted of 253 mammograms obtained in 103 patients. Each image contained a biopsy-proved mass that was evaluated in this study. Some cases involved multiple views or images from multiple examinations. The cases were randomly selected from patient files from the breast imaging division of a National Cancer Institute–designated national cancer center with the approval of the Institutional Review Board. The PPV of masses recommended for biopsy at this center is about 25%–30%, but an approximately equal number of malignant and benign masses (127 and 126, respectively) were chosen to enhance the statistical power in this observer performance study. Any images that were judged to be technically poor were excluded.

The mammograms were acquired with a contact technique. The dedicated mammographic systems had a molybdenum anode and molybdenum filter, a 0.3-mm nominal focal spot, and a reciprocating grid. MinR/MinR-E screen-film systems (Eastman-Kodak, Rochester, NY) were used with these units. Sixty-two of the malignant masses and six of the benign masses were judged to be spiculated by a radiologist (M.A.H.) experienced in mammography. The radiologist also measured the size (ie, longest dimension) and ranked the visibility of the masses on a scale of 1 (obvious) to 5 (subtle) relative to the range of visibility of masses encountered in clinical practice. For a description of the masses included in the data set, histograms of the size and visibility of the masses are shown in Figures 1a and 1b, respectively.

For the computer analysis, the selected mammograms were digitized with a laser



a. b. c.

d.

Figure 2. Example of rubber-band-straightening transform for extraction of texture features in the margin region surrounding a mass. **(a)** Original and **(b)** background-corrected images showing the region of interest with the mass, **(c)** mammogram showing an outline of the segmented mass, and **(d)** rubber-band-straightening–transformed image of a 40-pixel-wide region surrounding the segmented mass.

imager (Lumisys DIS-1000, Los Altos, Calif) at a pixel size of 0.1 × 0.1 mm and 12-bit gray levels. This imager has an optical density range of about 0.0–3.5. The optical density on the film was digitized linearly to pixel value at a calibration of 0.001 optical density unit/pixel value in the optical density range of about 0.0–2.8. The digitizer deviated from a linear response at an optical density higher than 2.8.

For the observer experiments, we used laser-printed images of the digitized mammograms for all readings. The images were printed with a 969HQ laser imager (Imation, Oakdale, Minn) that was connected to a Macintosh computer (Apple Computer, Cupertino, Calif) through a special digital interface. The interface provided a 12-bit in, 10-bit out look-up table and allowed images to be scaled to different factors with 15 interpolation methods. Because this laser imager has a pixel size of about 0.085 mm, we enlarged the images by about 18% during printing to maintain them at the same size as the original mammograms. One of the interpolation methods was chosen by an experienced radiologist (M.A.H.), who inspected the printed images with a magnifier and evaluated the sharpness of the spicules and mass boundaries. Because of the small pixel size used for both digitization and printing, basically no noticeable blurring of the masses could be seen with the chosen interpolation method. The images were also inspected for the potential contouring effect of 10-bit output images, but no noticeable artifacts could be found. A linear pixel value–to–output optical density calibration curve of the laser imager was used for the printing. All images were printed with the same settings.

Computerized Classification of Masses

Our computerized method of classifying mammographic masses has been described in detail previously (15–17). The method is summarized as follows: A region of interest that contained the biopsy-proved mass was identified on the mammogram by the radiologist. Background correction based on a distance-weighted estimation method was applied to the region of interest to reduce the low-frequency density variation in the region. A median-filtered smoothed image and two high-frequency enhanced images were generated from the background-corrected region of interest. The smoothed and enhanced gray-level values at each pixel were used as features in a k-means clustering algorithm to classify the pixels into two clusters; one was the mass, and the other was the surrounding breast tissue background. By choosing an appropriate criterion, a mass region slightly smaller than the actual mass that was visible on the image was segmented.

The boundary of the segmented region was smoothed by morphologic filtering. A new image transformation technique, referred to as the rubber-band-straightening transform, was used to transform a 40-pixel-wide region that surrounded the segmented mass boundary into a rectangular region. After transformation, the mass margin became approximately parallel, and any spicules that were radiating from the mass became approximately perpendicular, to the long dimension of the rectangular region. The rubber-band-straightening transform enabled the spicules to be aligned approximately in a uniform direction and thus facilitated the extraction of texture features from the margin of the mass. An example of a rubber-band-straightening–transformed image is shown in Figure 2.

Two types of texture features were found to be useful for classification. The first set of features included eight texture measures derived from the spatial gray-level dependence matrices of the rubber-band-straightening–transformed image. A spatial gray-level dependence matrix ele-

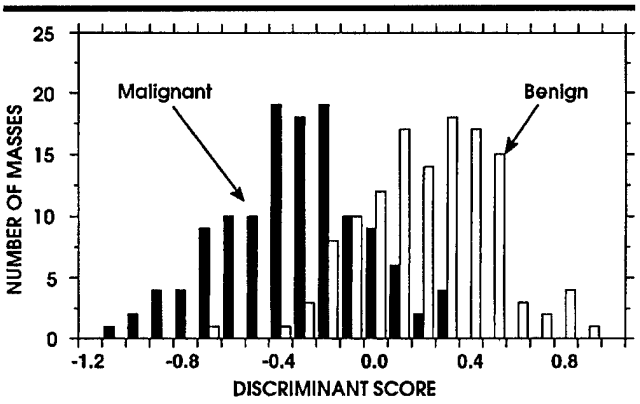


Figure 3. Histogram of the test discriminant scores of the 253 masses obtained from the linear discriminant classifier by using a "leave one case out" training and test resampling scheme. For this classifier, a smaller discriminant score corresponded to a higher likelihood of malignancy. The discriminant scores were used as the decision variable in the ROC analysis of classification performance.

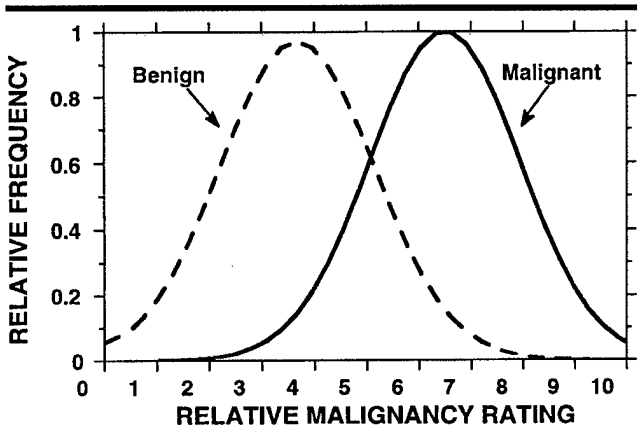


Figure 4. Binormal distribution fitted to the histogram of the discriminant scores of the malignant and benign masses. The discriminant scores were linearly transformed into a relative malignancy rating ranging from 1 to 10, where 1 corresponded to the most benign rating and 10 corresponded to the most malignant rating. This binormal distribution was shown to the observers during the training session to explain the rating scale of the computer classifier.

ment $p_{\theta,d}(i,j)$ is the joint probability of the occurrence of gray levels i and j for pixel pairs that are separated by a distance d and at a direction θ (22). For analysis of the masses, the spatial gray-level dependence matrices were constructed for 10 pixel distances (d = 1, 2, 3, 4, 6, 8, 10, 12, 16, 20 pixels) and in four directions (0°, 45°, 90°, 135°) relative to the mass boundary. Therefore, a total of 320 spatial gray-level dependence texture features were extracted.

The second set of texture features was derived from the run length statistics matrices of the horizontal and vertical gradient images of the rubber-band-straightening-transformed margin region. Five texture measures were extracted from the run length statistics matrix in each of the two directions (0° or 90°) on each gradient image. A total of 20 run length statistics texture features were thus obtained. Therefore, we had a total of 340 features from the two types of texture measures.

A stepwise linear discriminant feature selection procedure (23) was used to select the most effective features from the available feature set. A total of 41 features were selected. The selected features were input into the Fischer linear discriminant classifier (24) as predictor variables. A "leave one case out" resampling scheme was used to train and test the classifier. A histogram illustrating the test discriminant scores of the 253 masses is shown in Figure 3. For this classifier, a smaller discriminant score corresponded to a higher likelihood of malignancy. By using the test discriminant score as the decision variable, the performance of the computer classifier could be evaluated by using ROC analysis (17,25,26) and compared with that of the radiologists, as described later.

Relative Malignancy Rating of the Masses

For the observer performance study, we provided a relative malignancy rating of each mass to the observer during the reading session with CAD. The relative malignancy rating was obtained by taking a linear transformation of the computer classifier's decision variable to a range of 1–10 and rounding the value to the nearest integer. The transformation also reversed the relative magnitude of the decision variables so that 1 corresponded to the highest benignity rating, and 10 corresponded to the highest malignancy rating.

The purpose of the transformation was to provide a simple and intuitive relative scale for the observer. Because the transformation was linear and monotonic, the distributions of the normal and abnormal samples, as well as their ROC curves, were not affected, with the exception of a small error caused by making the decision variables discrete. Furthermore, the slope a and intercept b parameters that were fitted to the transformed discriminant scores for the normal and abnormal samples by using the LABROC program (26) were used to generate a binormal distribution. The fitted binormal distribution with the relative malignancy rating on a 1–10 scale (Fig 4), together with the computer's ROC curve, were shown and explained to the observers during a training session.

Observer Performance Study

Two ROC experiments (27) were conducted: The masses were evaluated from a single view in the first experiment and from two views in the second experiment. The location of the biopsy-proved mass was marked on each image so that the correct mass was evaluated by all observers. The observers were instructed to ignore any other possible masses on the images. Six radiologists (M.A.H., M.A.R., T.E.W., D.D.A., C.P., J.S.N.) who are approved by the Mammography Quality Standards Act and have 7–20 years of experience in interpreting mammograms participated in the observer performance experiments.

There were two reading sessions in each experiment—one with CAD and the other without CAD. The observers were asked to rate the likelihood of malignancy of the masses on a 10-point confidence rating scale under all reading conditions. In the first session, half the observers interpreted the images without CAD, and the other half interpreted them with CAD. The two reading sessions in the same experiment were separated by at least 3 weeks, and the two experiments were separated by 6 months. For all four reading sessions, the observer had unlimited time to read each case. To estimate the average reading time per case for each observer, the reading time for each case was recorded by using a stopwatch.

In the first experiment, the data set of 253 single-view mammograms was divided into a training set of 15 mammograms and a study set of 238 mammo-

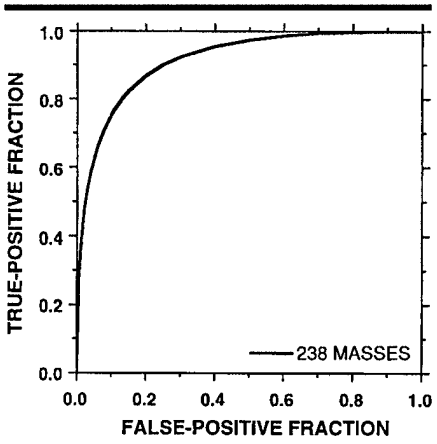


Figure 5. ROC curve for computerized classification of the 238 masses used in the observer performance study with single-view reading. The computer's ROC curve can be compared with the radiologists' ROC curves obtained from the single-view reading experiment illustrated in Figures 6 and 8.

grams (117 benign, 121 malignant). In each reading session, training was conducted before the reading of the study images. For the reading session with CAD, the fitted binormal distributions of the computer rating scores (Fig 4) for the entire data set were explained to the observer during training to familiarize the observer with the computer's rating scale. The computer rating of the mass was displayed on each image. After reading each training image, the observer was told the results of biopsy of the mass.

Each observer read the entire data set in one reading session. The order of the study images was randomized by a random number generator. The random sequence was different for each observer and for each reading session by the same observer. For the reading session with CAD, the observer was free to look at the computer rating, which was displayed on the image, either before or after estimating the likelihood of malignancy of the mass. However, each observer was asked to always read the computer rating before making a final decision. The observer was not informed of the pathologic results of any mass on the study images.

The second experiment was very similar to the first experiment. From the 238 single-view mammograms, 76 matched pairs (37 benign, 39 malignant) of craniocaudal and mediolateral oblique or lateral views were found. Another six pairs of two-view mammograms were identified from the rest of the images and used as training cases. The remaining mammograms were either single-view images or additional views of the pairs already chosen, so they were not used in this experiment. In this experiment, the observers were not informed of the pathologic results of any study case in any reading session. The 76 pairs of mammograms were read in one reading session by each observer.

For the reading session with CAD, the rating of the mass in each view was displayed on the respective image. The computer ratings of the mass on the two views were generally different. It was up to the observer to decide how to merge the two-view information. Observers were asked to give a single rating of the mass after reading both views.

ROC Analysis

The confidence ratings of each observer obtained from each reading condition were analyzed by using ROC methodology, and the classification accuracy was quantified by using the area under the ROC curve, A_z. A maximum likelihood estimation of the binormal distribution was fitted to the confidence ratings by using the LABROC program. This program provides an estimate of the A_z and of the a and b parameters of the ROC curve. The statistical significance of the difference in A_z between the reading with CAD and that without CAD was estimated with two methods: One was the Student paired t test for observer-specific paired data; the other was the Dorfman-Berbaum-Metz method for analysis of multireader, multicase ROC data (28). The statistical significance of the difference in A_z for reading single-view and two-view mammograms was estimated by using the Student paired t test for the six observers. The Student paired t test takes into account the statistical variation of readers, whereas the Dorfman-Berbaum-Metz method considers both reader variation and case sample variation by means of an analysis of variance approach. Therefore, the results of Dorfman-Berbaum-Metz analysis can be generalized to the population of readers as well as to the population of case samples.

Positive Predictive Value

An ROC curve represents the entire range of operating conditions of a diagnostic process and is independent of disease prevalence. When the disease prevalence is known, any operating point on an ROC curve can be used to derive the PPV and the corresponding false-negative fraction (false-negative fraction = 1 − true-positive fraction) on the basis of the following relationship: $PPV = TPF \times P(M) / [TPF \times P(M) + FPF \times P(B)]$, where TPF is the true-positive fraction, FPF is the false-positive fraction at the chosen decision threshold, and P(M) and P(B) are the prevalences of malignant and benign cases, respectively. By varying the decision threshold, the dependence of the PPV on the false-negative fraction can be derived.

Because our data set did not include masses on which biopsy had not been performed, the ROC curves obtained in this study cannot be generalized to predict the performance of the computer classifier and radiologists in clinical practice. However, to demonstrate the possible effect of CAD on the PPV in the population of masses in which biopsy is likely to be performed under the current clinical criteria, we can estimate the PPV by using the prevalence of the malignant and benign masses in this patient group. Because the PPV of masses sent for biopsy ranges from about 25% to 44% in general and from about 25% to 30% at our institution, for the purposes of our estimation, we assumed that the P(M) was 25% and the P(B) was 75% in this population. A higher prevalence of malignant cases would cause an increase in the PPV, but the trend between the PPV curves with and without CAD would be similar.

RESULTS

The ROC curve illustrating the performance of the computer classifier for the 238 study mammograms is shown in Figure 5. The ROC curve for the entire set of 253 mammograms (not shown) was almost identical to that of the 238 study cases; this indicates that the 15 training cases were typical of the 238 cases used in the study. The A_z values (± SD) for both ROC curves were 0.92 ± 0.02.

For the first experiment of reading the 238 single-view mammograms, the ROC curves for the readings by the six radiologists both without and with CAD are shown in Figures 6a and 6b, respectively. The A_z values of the six radiologists for the readings with and without CAD are listed in Table 1.

For the second experiment of reading the 76 pairs of two-view mammograms, the ROC curves for the readings by the six radiologists both without and with CAD are shown in Figures 7a and Figure 7b, respectively. The A_z values of the six radiologists in this experiment are also listed in Table 1.

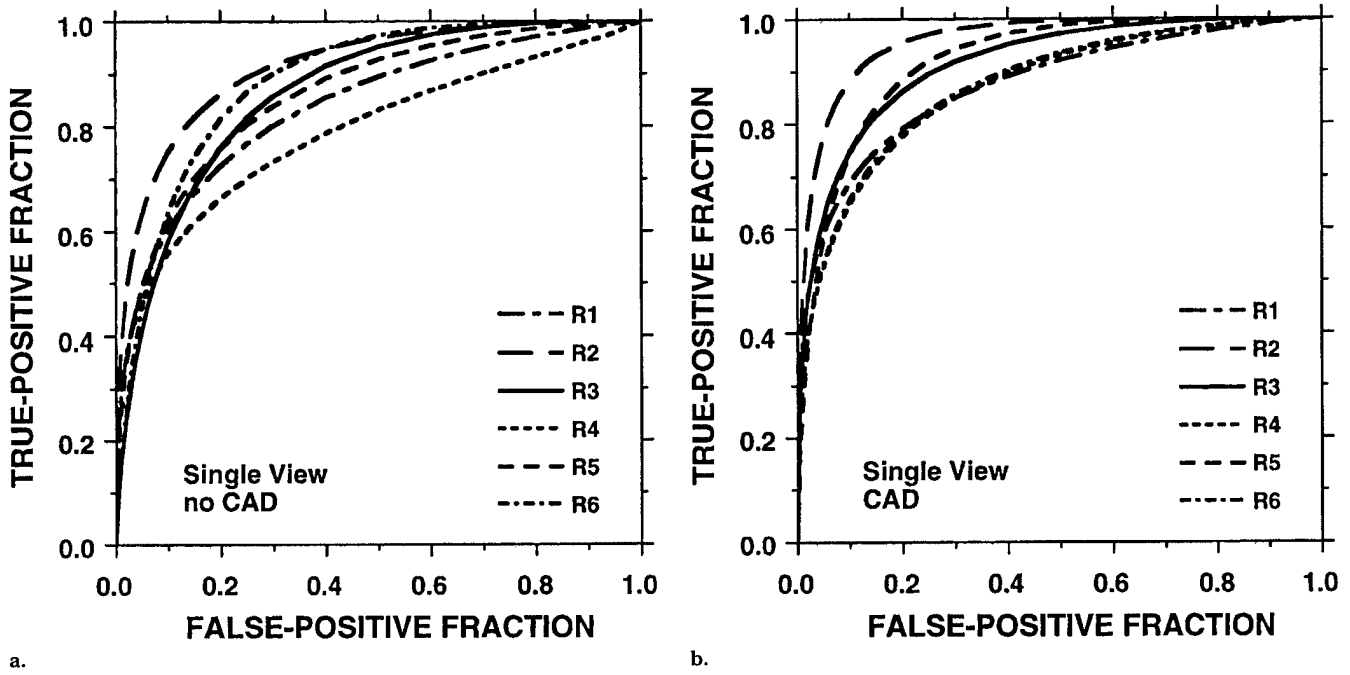


Figure 6. ROC curves for the six observers for single-view reading of the masses (a) without CAD and (b) with CAD. (a, b) *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6. Five of the six observers achieved an increase in the area under the ROC curve, A_z, with CAD.

The average ROC curve was derived from the average *a* and *b* parameters of the six individual ROC curves for a given reading condition (27). The average ROC curves for the four reading conditions are shown in Figure 8. The A_z values of the average ROC curves are listed in Table 1.

For the reading of the single-view mammograms, the performance of the computer classifier was comparable to that of the radiologist (reader 2) who had the highest classification accuracy (compare Figs 5 and 6) and higher than the average performance of the six radiologists (compare Figs 5 and 8). When the radiologists read the images with the computer aid, the classification accuracy of five radiologists improved (Table 1); the improvement in their A_z values ranged from 0.04 to 0.08. The average performance of the six radiologists became comparable to that of the computer classifier. The improvement in the radiologists' classification accuracy by using CAD was statistically significant ($P = .022$, Student paired *t* test; $P = .020$, Dorfman-Berbaum-Metz method). Reader 2 with CAD obtained an A_z value of 0.96, which was higher than that obtained by the radiologist alone or by the computer alone.

A trend similar to that with the single-view readings was observed with the two-view readings. The A_z value of the computer classifier for the corresponding 152 single-view masses was 0.91 ± 0.02. The classification accuracy of all six radiologists improved when they read the mammograms with the computer aid. The increase in the A_z values ranged from 0.01 to 0.07. The improvement was statistically significant ($P = .007$, Student paired *t* test; $P = .026$, Dorfman-Berbaum-Metz method). With CAD, two radiologists achieved an A_z value of 0.97, which was higher than that obtained by the radiologists alone or by the computer alone. These results indicate that the second opinion provided by the computer classifier might have strengthened the radiologists' confidence in the interpretation of some difficult cases but had less influence on the radiologists' decision when the computer made mistakes or when the radiologists were confident about their decision.

As can be seen from the data in Table 1,

TABLE 1
Areas under the ROC Curves for the Classification of Masses with and without CAD by the Six Radiologists

Radiologist No.	A_z (Single View)*		A_z (Two View)†	
	Without CAD	With CAD	Without CAD	With CAD
1	0.84 ± 0.03	0.87 ± 0.02	0.90 ± 0.03	0.93 ± 0.03
2	0.92 ± 0.02	0.96 ± 0.01	0.95 ± 0.02	0.97 ± 0.02
3	0.86 ± 0.02	0.91 ± 0.02	0.92 ± 0.03	0.93 ± 0.03
4	0.79 ± 0.03	0.87 ± 0.02	0.88 ± 0.04	0.95 ± 0.03
5	0.86 ± 0.02	0.92 ± 0.02	0.93 ± 0.03	0.97 ± 0.02
6	0.89 ± 0.02	0.87 ± 0.02	0.89 ± 0.04	0.93 ± 0.03
A_z from average a, b parameters	0.87	0.91	0.92	0.96

Note.—Data are the mean ± SD.

* $P = .022$ for the difference between the A_z values measured with CAD and those measured without CAD, as determined by using the Student two-tailed *t* test. $P = .020$ for this difference, as determined by using the Dorfman-Berbaum-Metz method.

† $P = .007$ for the difference between A_z values measured with CAD and those measured without CAD, as determined by using the Student two-tailed *t* test. $P = .026$ for this difference, as determined by using the Dorfman-Berbaum-Metz method.

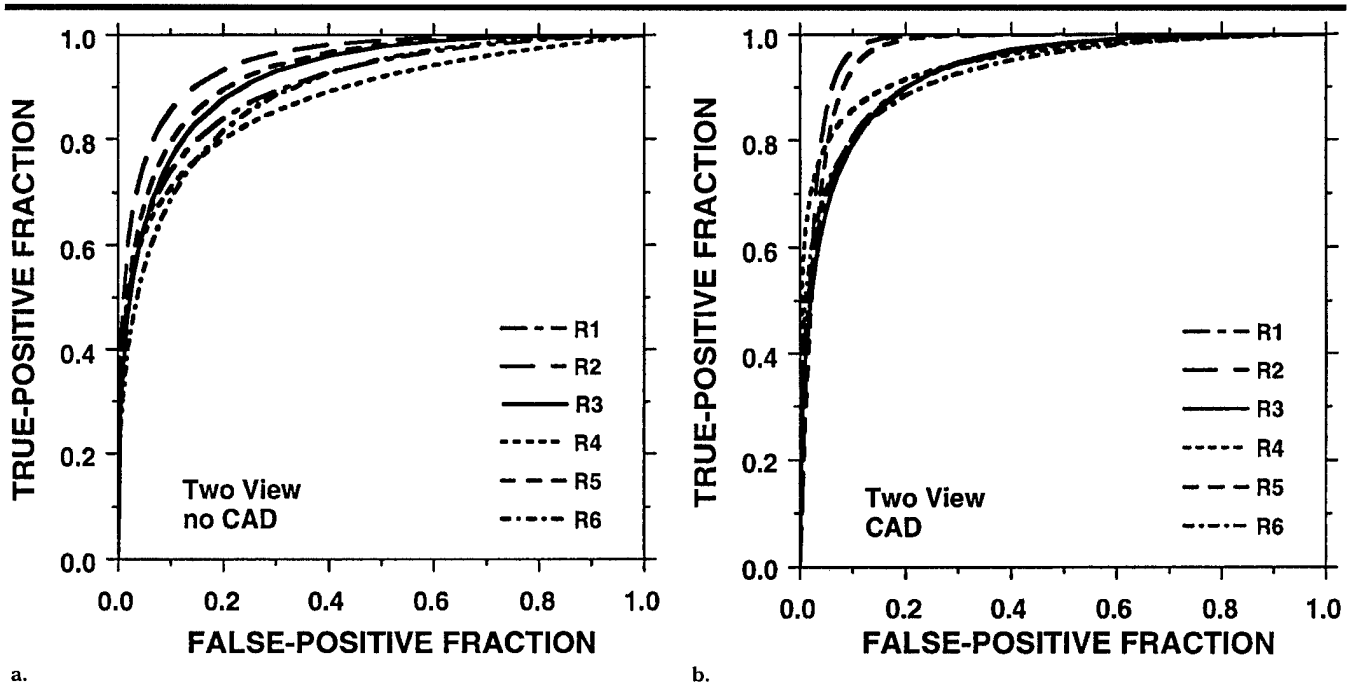


Figure 7. ROC curves for the six observers for two-view reading of the masses (a) without CAD and (b) with CAD. (a, b) *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6. All six observers achieved an increase in the area under the ROC curve, A_z, with CAD.

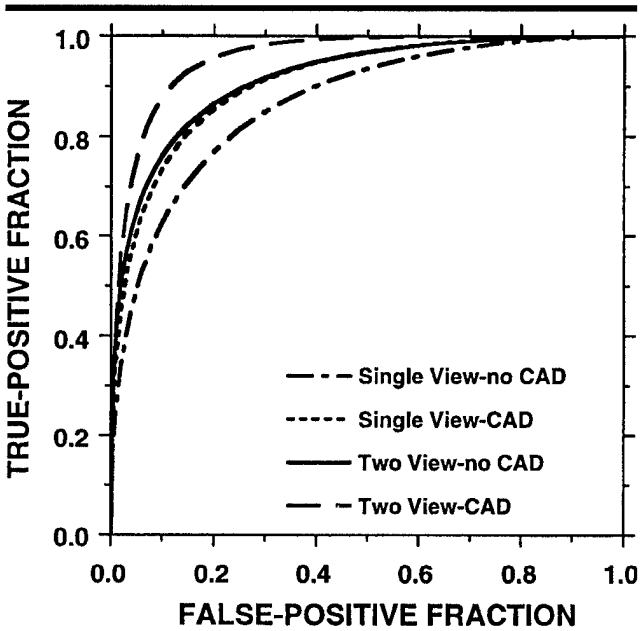


Figure 8. Average ROC curve obtained from the average *a* and *b* parameters of the six individual ROC curves for each of the four reading conditions. An improved ROC curve was achieved with CAD in both the single-view and two-view reading experiments.

the radiologists' accuracy in classifying masses by reading two-view mammograms was consistently higher than that by reading single-view mammograms (P = .008). This trend remained when they read the mammograms with CAD (P = .007). These findings are consistent with the clinical experience of the radiologists that at least two views of mammograms are needed to effectively evaluate a suspicious lesion.

The PPV as a function of the false-negative fraction was derived from the fitted ROC curves under the assumption that the prevalence of malignant masses was 25% in the population of masses sent for biopsy. The PPVs estimated for the six observers who read the two-view mammograms with and without CAD are plotted in Figure 9. CAD would provide an improvement in the PPV in the high false-negative fraction range for all observers except readers 2 and 5. The increase in the PPV at a decision threshold of "no missed malignant mass" (ie, false-negative fraction = 0) varied over a wide range; the largest gain, 39%, would be achieved by reader 2, and the smallest gain, 0%, would be achieved by reader 4.

DISCUSSION

In the observer experiment of reading two-view mammograms with CAD, we presented the computer's rating of each view separately. The decision of how to merge the computer ratings of the two views was left to the radiologist. It is likely that the radiologists took the conservative approach of using the highest malignancy rating of the two as the computer's overall rating. However, it also might have depended on whether the relative ranking between the two computer ratings agreed with the observer's opinion. In some cases, we observed that the radiologist's rating was very different from the computer's rating of either view.

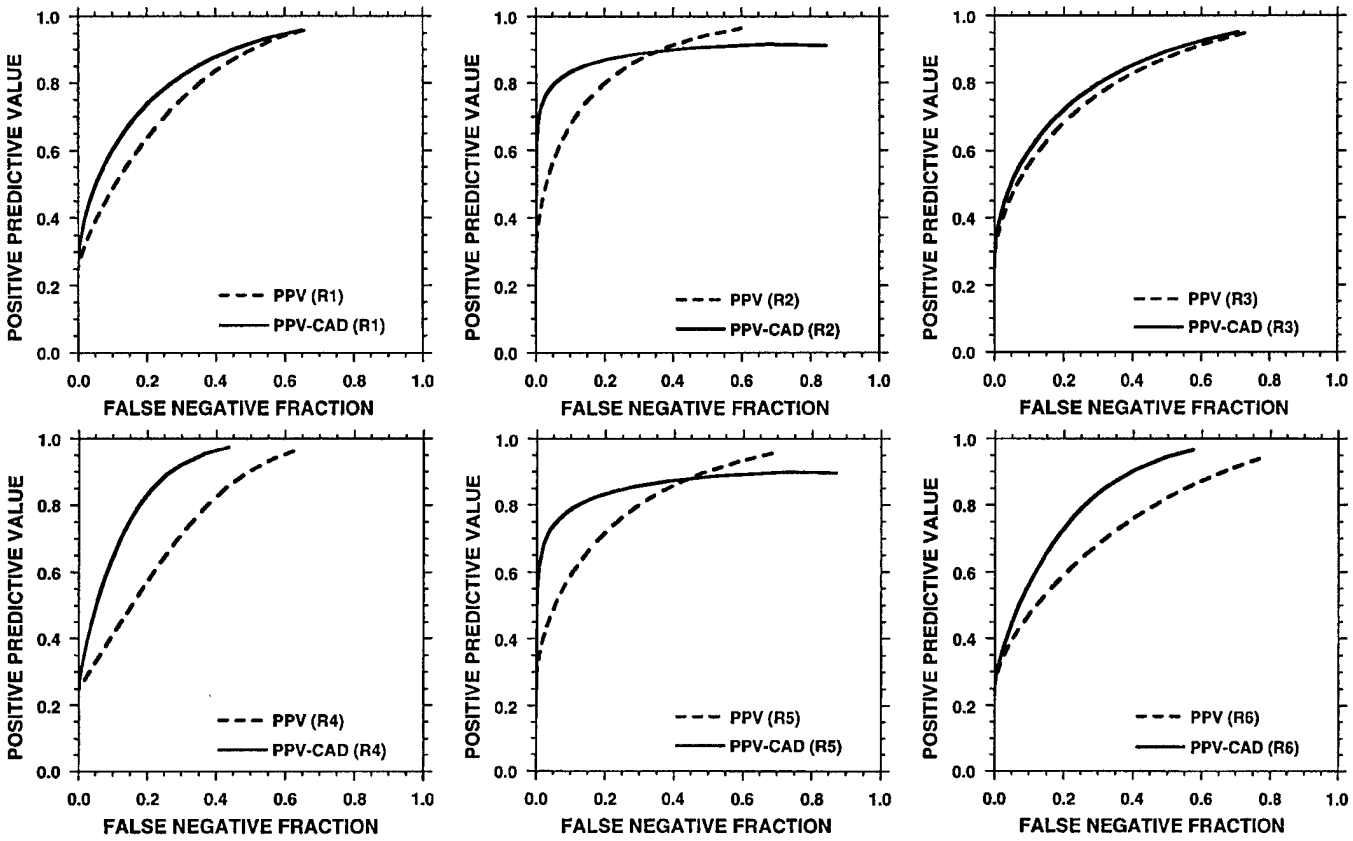


Figure 9. PPV as a function of the false-negative fraction derived from the ROC curves for the six observers (Fig 7). The PPV was predicted for a population of masses in which biopsy was likely to be performed under current clinical criteria and by assuming the prevalence of malignant masses to be 25%. *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6.

Because decision making is a complex process, the simple approach of using the highest malignant rating or the average rating from multiple views may not be the method preferred by radiologists. The separate ratings that we used in this study would provide less biased information. Further investigation is needed to determine the best approach of presenting the computer's ratings to radiologists in clinical practice.

To obtain insight into how the radiologists might use the two-view information, we compared the classification results from their true two-view reading with those from a simulated two-view reading without the computer aid. The latter results were derived from ratings of single-view readings of the same 76 pairs of mammograms interpreted in experiment 2 by assuming two strategies—one in which the highest malignancy rating between the two ratings was used, and the other in which the average of the two ratings was used (Table 2). The A_z values for these classification ratings derived from the single-view reading are listed in Table 2. The corresponding A_z values for the computer classifier are also given in Table 2 for comparison.

The A_z values for the maximal rating and the average rating were similar. Four of the radiologists obtained higher A_z values at the true two-view reading; the A_z values obtained by the remaining two radiologists were lower than those obtained at the simulated two-view reading. Although the difference did not achieve statistical significance ($P = .37$) and both readings included intraobserver variations, there seemed to be a slight trend toward the true two-view reading being more accurate than the simulated two-view reading. This may indicate that the radiologists used a more complex decision-making process to interpret the two views of the masses than that of simply maximizing or averaging the ratings from each view.

In this study, the discriminant scores of the masses given by the computer classifier were transformed into a relative malignancy rating. The relative malignancy rating scale and the distribution of the malignant and benign masses along the relative rating scale were explained to the observers in the training sessions. A relative malignancy rating scale was used because the true likelihood of malignancy of the masses could not be estimated from a small data set, as will be explained. However, the relative rating scale provided by the computer was ad-

TABLE 2
Estimation of the Malignancy Classification of 76 Masses by Two-View Reading, as Simulated from Single-View Reading of Mammograms by Radiologists without CAD

Radiologist No.	A_z	
	Maximal Rating	Average Rating
1	0.94 ± 0.03	0.93 ± 0.03
2	0.94 ± 0.03	0.94 ± 0.03
3	0.84 ± 0.05	0.86 ± 0.04
4	0.85 ± 0.04	0.83 ± 0.05
5	0.88 ± 0.04	0.89 ± 0.04
6	0.91 ± 0.03	0.92 ± 0.03
Computer	0.96 ± 0.02	0.96 ± 0.02

Note.—Data are the mean ± SD. Two strategies were used: In one, the highest of the malignancy ratings on each view was used; in the other, the average between the two ratings was used.

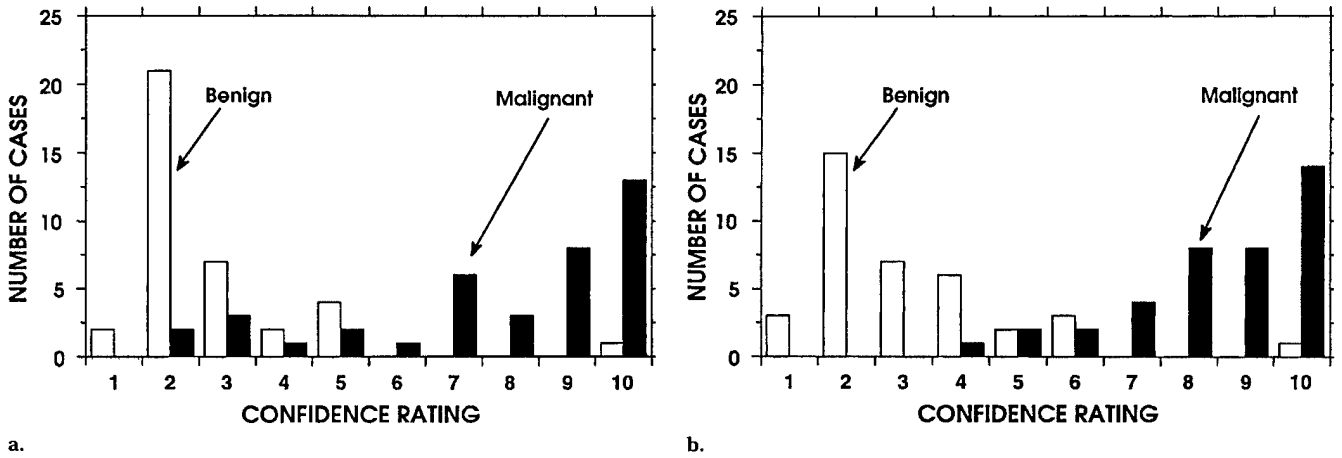


Figure 10. Histograms illustrate the confidence ratings of reader 5 obtained by reading 76 two-view mammograms **(a)** without CAD and **(b)** with CAD. The specificity of reader 5 at 100% sensitivity would increase from 5% (two of 37 masses) without CAD to 68% (25 of 37 masses) with CAD if an appropriate decision threshold were chosen.

equate for measuring the relative performance of classification with and without CAD in an ROC study.

If a computer classifier is trained and tested with very large data sets, and if both the malignant and benign cases represent random samples of the population, then the likelihood of malignancy of a classified mass can be estimated on the basis of the probability distributions of the classifier's test output scores and the prevalence of the two classes of masses in the patient population. However, with a relatively small data set, such as that used in this and other observer studies (14), there are limitations. First, the performance of a classifier trained with a small sample set may have large bias and variance (29–31). Second, the data set in this study did not include masses on which biopsy was not performed, so it did not represent a random sample of the masses in the patient population. If our classifier were applied to all cases of solid masses in clinical practice, the probability distribution of the test scores for the two classes of masses would be different from that of the current data set.

If we ignore the patient population at large, it is possible to estimate the likelihood of malignancy of a mass on the basis of the probability distribution of the classifier output scores by using the prevalence of the two classes of masses in this specific data set. However, the likelihood of malignancy derived in this way will be completely different from the true likelihood of malignancy of a mass in the patient population. This can be easily seen if one considers that the same mass with the same discriminant score will have a smaller likelihood of malignancy if it is analyzed within a data set that has a lower prevalence of malignant cases than that in the current data set.

Training the participating radiologists with a "likelihood of malignancy" derived from a small data set for the observer experiment may mislead them if they encounter a similar mass in their clinical practice. We, therefore, preferred to use a "relative malignancy rating," which is independent of the prevalences of malignant and benign masses in the data set. As long as the same classifier and the same linear transformation are used for classifying masses, the relative malignancy rating for a given mass will remain the same, regardless of the types of other masses in the data set. When a computer classifier is implemented in a clinical setting and its performance can be established in the patient population, the true likelihood of malignancy of a given mass can be estimated and provided to the radiologist. The true likelihood of malignancy may be a more informative measure for radiologists in the clinical application of CAD.

For the reading of the 76 two-view mammograms, the results of the ROC study indicated an improvement in the A_z value for all six radiologists when the computer aid was used. This indicates an overall increase in the separation of confidence rating distributions between the malignant and benign cases. The histograms in Figure 10 illustrate the distributions of confidence ratings with and without CAD for reader 5, who achieved the second greatest improvement in both the A_z value (Table 1) and the separation of malignant from benign distributions. Without CAD, this reader's ratings of the malignant cases ranged from 2 to 10. This is consistent with the fact that biopsy was performed in all masses in the data set to avoid missing the malignant cases. With CAD, there was marked improvement in the separation of the two distributions. It is possible to set a decision threshold at a confidence rating of 4, below which biopsy would not need to be performed and no malignant masses would be missed. The number of benign masses that could be identified without missing a malignant mass by setting an appropriate threshold would increase by 23 (out of 76 cases) for reader 5. Five of the six radiologists in our ROC study achieved an improvement in distinguishing benign from malignant masses, and one radiologist had no difference. Although the improvement of the five radiologists varied over a wide range, from one to 25 cases, this result indicates a strong possibility that CAD can be used to reduce the number of unnecessary biopsies.

The large variation in improvement among the radiologists may have been due to the different degrees of confidence that they had in the computer aid. As with any new diagnostic tool, this confidence is influenced by the experience the radiologist has with the tool. Although the radiologists received training before the reading sessions, the high variability in confidence was not unexpected, because this ROC study was the first instance in which they had worked with the computer aid. Their confidence levels may have also been reflected in the relatively low accuracy of classification by some radiologists with CAD compared with that of the computer classifier alone.

If a radiologist can increase his or her

confidence in the performance of a computer aid by gaining more extensive clinical experience, then he or she will likely be able to find the most effective way of merging his or her judgment with the computer's rating and thus reduce both interobserver and intraobserver variability. Because a radiologist who uses CAD can establish a meaningful decision threshold for biopsy only after becoming familiar with the sensitivity and specificity of working with CAD, the radiologists in this study were not asked to decide whether biopsy should have been performed on a mass. Rather, we focused on the evaluation of changes in the sensitivity and specificity of the radiologists' classification of masses when CAD was used.

In this ROC study, all six observers were attending radiologists with extensive experience in the interpretation of mammograms. It is possible that the computer aid may be even more useful to radiology residents or radiologists with less experience in mammography. The effect of CAD on mammographic interpretation by less-experienced readers will be a subject of investigation in future studies.

The observers were allowed unlimited time to read each case in this ROC study. To obtain an estimate of the change in reading time with CAD, we recorded the reading time of each observer in each reading session by using a stopwatch. For the single-view reading experiment, the average reading time per image without CAD varied from 4.3 seconds to 17.1 seconds (mean time for the six observers, 7.8 seconds). The average reading time per image with CAD varied from 4.2 seconds to 17.3 seconds (mean time, 7.3 seconds). For the two-view reading experiment, the average reading time per pair of images without CAD varied from 6.6 seconds to 16.0 seconds (mean time, 10.4 seconds). The average reading time per pair of images with CAD varied from 7.6 seconds to 27.1 seconds (mean time, 13.5 seconds).

The reading time essentially did not change with use of the computer aid for the single-view readings. For the two-view readings, the radiologists took longer with CAD, probably because they had to merge the two computer ratings and merge the computer ratings with their own evaluations. Further investigation is needed to determine whether there is a trade-off between the radiologist's efficiency and the method of presenting the computer rating and whether the reading time with CAD will depend on the experience that the radiologist has with the computer information.

In the observer study, we used laser-printed mammograms instead of the original mammograms for the reading experiments. A major reason is that it is difficult to keep all the original mammograms together for the entire period of the study because they are part of active patient files and thus often recalled for comparison with new studies or for other clinical reasons. Because the maximum optical density of laser-printed images was 3.1 for the laser imager used, the contrast on the printed mammograms was about 20% lower than that on the original mammograms. Although the image quality was slightly lower than that of the original, the laser-printed digitized images were judged to be adequate for reading the details of the masses by the participating radiologists. The laser-printed image set might also be considered as one that had slightly more subtle masses than the original set of images. Because the relative performance of two modalities is measured in ROC experiments, and because the readings both with and without CAD in this study were conducted with the same set of printed images, the relative performance of the two readings should be valid. It should also be noted that in order for a computer aid that uses automated image analysis to be widely accepted, direct digital mammography would have to be the imaging modality in clinical use. Laser-printed images or soft-copy monitors will be the display medium for the digital mammograms. The use of laser-printed images for this ROC study was therefore practical.

In our observer performance experiment, we found that CAD improved the radiologists' ability to distinguish malignant and benign masses. This is consistent with the results of other studies (11,14) in which a statistically significant improvement ($P < .001$ in both studies) in the radiologists' classification accuracy by using CAD was found. The results of the former study (11) further showed that the PPV of a recommendation for biopsy by the radiologists was significantly increased ($P < .001$). In our approach, the computer classifier automatically extracted image features, whereas in the other studies, the computer classifier used the radiologist's evaluation and other patient information as input. Therefore, it appears that CAD can provide a useful second opinion to radiologists, either by consistently extracting and analyzing the image features or by optimally weighting various diagnostic factors and thereby improving the consistency in the decision-making process. This suggests that a computer classifier that combines both approaches—that is, automatically extracts image features and optimally merges them with the radiologist's evaluation and patient information—may be even more effective for breast cancer diagnosis. The latter step will also improve the radiologist's utilization of the computer rating on the basis of the computer-extracted features; this utilization was found to have large interobserver variation in our ROC experiment.

In conclusion, an ROC study of the effects of CAD on radiologists' classification of malignant and benign masses on mammograms was conducted. The results showed that CAD can provide a statistically significant improvement in the classification accuracy—that is, in the A_z value—for both single-view reading ($P = .022$) and two-view reading ($P = .007$). The improved separation between the confidence ratings of the malignant masses and those of the benign masses indicates the potential that CAD may reduce the rate of biopsy of benign masses when decision thresholds are properly chosen by the radiologists. The decision threshold may vary among radiologists, as in the case of mammographic interpretation without CAD, and can be set after the radiologist working with CAD has established his or her sensitivity and specificity with this approach through clinical experience.

Further studies are needed to evaluate the effects of CAD on the accuracy of radiologist classification of masses in clinical settings in which the prevalence of malignant masses is different from that in a laboratory data set and the likelihood of malignancy of a mass can be estimated by the computer classifier. In the two-view reading ROC experiment, the reading time per case increased by about 30% with the use of CAD. The dependence of the radiologist's efficiency in reading with CAD on the presentation method and on the reader's experience in using the computer information also warrants further investigation.

Acknowledgments: The authors are grateful to Charles E. Metz, PhD for useful discussions and for the use of the LABROC and LABMRMC programs.

References

1. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics 1998. CA Cancer J Clin 1998; 48:6–29.
2. Adler DD, Helvie MA. Mammographic biopsy recommendations. Curr Opin Radiol 1992; 4:123–129.

3. Kopans DB. The positive predictive value of mammography. AJR 1991; 158:521–526.
4. Shtern F. Digital mammography and related technologies: a perspective from the National Cancer Institute. Radiology 1992; 183: 629–630.
5. Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
6. Vyborny CJ. Can computers help radiologists read mammograms? Radiology 1994; 191:315–317.
7. Chan HP, Doi K, Vyborny CJ, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990; 25:1102–1110.
8. Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW, Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191: 331–337.
9. Getty DJ, Pickett RM, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988; 23:240–252.
10. D'Orsi CJ, Getty DJ, Swets JA, Pickett RM, Seltzer SE, McNeil BJ. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. Radiology 1992; 184:619–622.
11. Baker JA, Kornguth PJ, Lo JY, Floyd CE. Artificial neural network: improving the quality of breast biopsy recommendations. Radiology 1996; 198:131–135.
12. Chan HP, Sahiner B, Petrick N, et al. Observer performance study of radiologists' reading of mammographic masses and comparison with computerized classification (abstr). Radiology 1996; 201(P):370.
13. Chan HP, Sahiner B, Helvie MA, et al. Effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses on mammograms: an ROC study (abstr). Radiology 1997; 205(P):275.
14. Jiang Y, Nishikawa R, Schmidt RA, Metz CE, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis (CAD): an observer study (abstr). Radiology 1997; 205(P):274.
15. Sahiner B, Chan HP, Petrick N, Helvie MA, Adler DD, Goodsitt MM. Classification of masses on mammograms using rubberband straightening transform and feature analysis. Proc SPIE 1996; 2710:44–50.
16. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: the rubber-band straightening transform and texture analysis. Med Phys 1998; 25:516–526.
17. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. Phys Med Biol 1998; 43:2853–2871.
18. Ackerman LV, Gose EE. Breast lesion classification by computer and xeroradiograph. Cancer 1972; 30:1025–1035.
19. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computerized image analysis. IEEE Trans Med Imaging 1993; 12:664–669.
20. Pohlman S, Powell KA, Obuchowshi NA, Chilote WA, Grundfest-Broniatowski S. Quantitative classification of breast tumors in digitized mammograms. Med Phys 1996; 23:1337–1345.
21. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol 1998; 5:155–168.
22. Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans Syst Man Cybernetics 1973; 3:610–621.
23. Norusis MJ. SPSS for Windows release 6: professional statistics. Chicago, Ill: Statistical Product for Service Solutions, 1993.
24. Lachenbruch PA. Discriminant analysis. New York, NY: Hafner, 1975; 8–19.
25. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
26. Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. Stat Med 1998; 17:1033–1053.
27. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
28. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol 1992; 27:723–731.
29. Fukunaga K, Hayes RR. Effects of sample size on classifier design. IEEE Trans Pattern Analysis and Machine Intelligence 1989; 11:873–885.
30. Chan HP, Sahiner B, Wagner RF, Petrick N, Mossoba J. Effects of sample size on classifier design: quadratic and neural network classifiers. Proc SPIE 1997; 3034:1102–1113.
31. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis in mammography: effects of finite sample size. Med Phys 1997; 24:1034–1035.

Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach

Lubomir Hadjiiski,* *Member, IEEE*, Berkman Sahiner, *Member, IEEE*, Heang-Ping Chan, Nicholas Petrick, *Member, IEEE*, and Mark Helvie

***Abstract*—A new type of classifier combining an unsupervised and a supervised model was designed and applied to classification of malignant and benign masses on mammograms. The unsupervised model was based on an adaptive resonance theory (ART2) network which clustered the masses into a number of separate classes. The classes were divided into two types: one containing only malignant masses and the other containing a mix of malignant and benign masses. The masses from the malignant classes were classified by ART2. The masses from the mixed classes were input to a supervised linear discriminant classifier (LDA). In this way, some malignant masses were separated and classified by ART2 and the less distinguishable benign and malignant masses were classified by LDA. For the evaluation of classifier performance, 348 regions of interest (ROI's) containing biopsy proven masses (169 benign and 179 malignant) were used. Ten different partitions of training and test groups were randomly generated using an average of 73% of ROI's for training and 27% for testing. Classifier design, including feature selection and weight optimization, was performed with the training group. The test group was kept independent of the training group. The performance of the hybrid classifier was compared to that of an LDA classifier alone and a backpropagation neural network (BPN). Receiver operating characteristics (ROC) analysis was used to evaluate the accuracy of the classifiers. The average area under the ROC curve (A_z) for the hybrid classifier was 0.81 as compared to 0.78 for the LDA and 0.80 for the BPN. The partial areas above a true positive fraction of 0.9 were 0.34, 0.27 and 0.31 for the hybrid, the LDA and the BPN classifier, respectively. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.**

***Index Terms*— Computer-aided diagnosis, hybrid classifier, mammography, neural networks.**

I. Introduction

MAMMOGRAPHY is the most effective method for detection of early breast cancer [1]. However, the specificity for classification of malignant and benign lesions from mammographic images is relatively low. Clinical studies have shown that the positive predictive value (i.e., ratio of the number of breast cancers found to the total number of biopsies) is only 15% to 30% [2]–[4]. It is important to increase the positive predictive value without reducing the sensitivity of breast cancer detection. Computer-aided diagnosis (CAD) has the potential to increase the diagnostic accuracy by reducing the false-negative rate while increasing the positive predictive values of mammographic abnormalities.

Classifier design is an important step in the development of a CAD system. A classifier has to be able to merge the available input feature information and make a correct evaluation. Commonly used classifiers for CAD include linear discriminants (LDA) [5], [6] and backpropagation neural networks (BPN) [7]–[9] which have been shown to perform well in lesion classification problems [10]–[22]. These classifiers are generally designed by supervised training. However, these types of classifiers have limitations dealing with the nonlinearities in the data (in case of LDA) and in generalizability when a limited number of training samples are available (especially BPN). Another classification approach is based on unsupervised classifiers, which cluster the data into different classes based on the similarities in the properties of the input feature vectors. Therefore, unsupervised classifiers can be used to analyze the similarities within the data. However, it is difficult to use them as a discriminatory classifier [29], [30]. They also have limited generalizability when the training sample set is small.

We propose here a hybrid unsupervised/supervised structure to improve classification performance. The design of this structure was inspired by neural information processing principles such as self organization, decentralization and generalization. It combines the adaptive resonance theory network (ART2) [26], [27] and the LDA classifier as a cascade system (ART2LDA). The self-organizing unsupervised ART2 network automatically decomposes the input samples into classes with different properties. The ART2 network has been found to perform better compared to conventional clustering techniques in terms of learning speed and discriminatory resolution for the detection of rare events in many classification tasks [28]–[30]. The supervised LDA then classifies the samples belonging to a subset of classes that have greater similarities. By improving the homogeneity of the samples, the classifier designed for the subset of classes may be more robust.

The ART2LDA design implements both structural and data decomposition. Decomposition is a powerful approach that can reduce the complexity of a problem. Both structural decom-

Manuscript received January 27, 1999; revised October 26, 1999. This work was supported by in part by the USPHS under Grant No. CA 48129 and in part by the U.S. Army Medical Research and Materiel Command (USAMRMC) under Grant DAMD 17-96-1-6254. The work of L. Hadjiiski was supported in part by the USAMRMC under Career Development Award DAMD 17-98-1-8211. The work of B. Sahiner was supported in part by the USAMRMC under Career Development Award DAMD 17-96-1-6012. The work of Nicholas Petrick was supported in part by a grant from The Whitaker Foundation. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author.*

*L. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, and M. Helvie are with the Department of Radiology, The University of Michigan, Ann Arbor, MI 48109-0904 USA.

Publisher Item Identifier S 0278-0062(99)10410-5.

position and data decomposition can improve classification accuracy [23] as well as model accuracy [24]. However, decomposition can also reduce the prediction accuracy due to overfitting the training data. We will demonstrate in this paper that the proposed hybrid structure can reduce the overfitting problem and improve the prediction capabilities of the system. The performance of the hybrid ART2LDA classifier will be compared with those of an LDA alone or a BPN classifier.

The rest of the paper is organized as follows. In Section II the ART2 unsupervised network is described. A hybrid ART2LDA classifier is introduced in Section III. Section IV describes the data set used in this study. The results are presented in Section V. Section VI contains discussion of these results. Finally, Section VII concludes this investigation.

II. ART2 Unsupervised Neural Network

The ART2 is a self-organizing system that can simulate human pattern recognition. ART2 was first described by Grossberg [25] and a series of further improvements were carried out by Carpenter, Grossberg, and coworkers [26]–[28]. The ART2 network clusters the data into different classes based on the properties of the input feature vectors. The members within a class have similar properties. The process of ART2 network learning is a balance between the plasticity and stability dilemma. Plasticity is the ability of the system to discover and remember important new feature patterns. Stability is the ability of the system to remain unchanged when already known feature patterns with noise are input to the system. The balance between plasticity and stability for the ART2 training algorithm allows fast learning [28], i.e., rare events can be memorized with a small number of training iterations without forgetting previous events. The more conventional training algorithms, such as back propagation [7]–[9], perform slow learning, i.e., they tend to average over occurrences of similar events and require many training iterations.

The structure of the ART2 system is shown in Fig. 1. It consists of two parts: the ART2 network and the learning stage. Suppose that there are n input features x_i $(i = 1, \cdots, n)$ and k classes in the ART2 network. When a new vector is presented to the input of the ART2 network, an activation value p_j for class j is calculated as

$$p_j = \sum_{i=1}^{n} x_i w_{ij}, \qquad j = 1, \cdots, k \tag{1}$$

where w_{ij} is the connection weight between input i and class j. The activation value is a measure of the membership of the particular input feature vector to class j. The higher the value p_j is, the better the input vector matches class j. The maximum value p_r is selected from all p_j $(j = 1, \cdots, k)$ to find the best class match. Furthermore, in order to balance the contribution to the activation value from all feature components, the input feature values applied to the ART2 system are scaled between zero and one [30]. This normalization will allow detection of similar feature patterns even when the magnitudes of the input feature components are very different.

The learning stage of the ART2 system can influence the weights of the selected class or the complete ART2 network

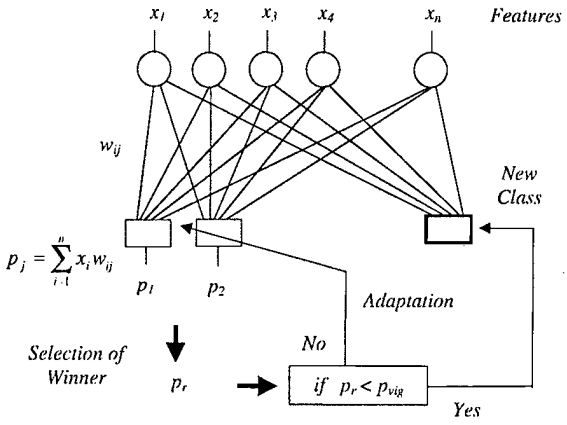


Fig. 1. Structure of the ART2 network.

structure by adding a new class. An additional parameter, the vigilance, is used to determine the type of learning [26]. The vigilance parameter p_{vig} is a threshold value that is compared to the maximum activation value p_r. If p_r is larger than p_{vig} then the input vector is considered to belong to class r. The adaptation of the weights connected with class r is performed as follows:

$$w_{ir}^{new} = w_{ir}^{\text{old}} + \eta(x_i - w_{ir}^{\text{old}}), \qquad \text{for } i = 1, \cdots, n \tag{2}$$

where η is a learning rate. The adaptation of the class r weights (2), aims at maximization of the p_r value for the particular input vector. In an iterative manner the weights are adjusted so that the activation values produced for similar input vectors will be maximum only for the class to which they belong and these maximum activation values will be higher than p_{vig}.

If the maximum activation value p_r is smaller than p_{vig}, it is an indication that a novelty has appeared and a new class will be added to the ART2 structure. The new weights connecting the input with the new class $(k + 1)$ are initialized with the scaled input feature values of this novelty. In such a way, the activation value p_{k+1} will be maximum $(p_r = p_{k+1})$ higher than p_{vig} when computed for this novelty in further training iterations. The value of the vigilance parameter p_{vig} determines the resolution of ART2. It can be chosen in the range between zero and one. In the case that p_{vig} is relatively small, only very different input feature vectors will be distinguished and separated in different classes. If p_{vig} is relatively large, the input feature vectors that are more similar will be separated into different classes. The value of p_{vig} is selected differently depending on the particular application.

III. ART2LDA Classifier

Despite the good performance of ART2 for efficient clustering and detection of novelties, the fast learning approach can cause problems associated with the generalization capability of the system and the correct classification of unknown cases. Supervised classifiers such as linear discriminants or backpropagation neural network classifiers can have better generalization capability than ART2, because they are trained by averaging over similar event occurrences. However, the learning process in these traditional learning algorithms tends

to erase the memory of previous expert knowledge when a new type of expertise is being learned. Therefore, these classifiers do not have as good an ability to correctly classify rare events as ART2 [28], [29].

In order to improve the accuracy and generalization of a classifier, we propose to design a hybrid classifier that combines the unsupervised ART2 network and a supervised LDA classifier. This hybrid classifier (ART2LDA) utilizes the good resolution capability of ART2 and the good generalization capability of LDA. The ART2 first analyzes the similarity of the sample population and identifies a subpopulation that may be separated from the main population. This will improve the performance of the second-stage LDA if the subpopulation causes the sample population to deviate from multivariate normal distributions for which LDA is an optimal classifier. Therefore, the ART2 serves as a screening tool to improve the homogeneity of the sample distributions by classifying outlying samples into separate classes.

The ART2LDA hybrid classifier can be described as

$$y_{AL} = g(f_2(x))f_1(x) + 1 - g(f_2(x)) \tag{3}$$

where x is the input vector, $f_1(\cdot)$ is the LDA classifier, $f_2(\cdot)$ is the ART2 classifier, and $g(\cdot)$ is a binary membership function, which labels the classes identified by ART2 to be one of the two types: malignant class or mixed class. A particular class is defined as malignant if it contains only malignant members. It is defined as mixed if it contains both malignant and benign members. The membership function is defined as follows:

$$g(c) = \begin{cases} 0, & \text{if } c \text{ is a malignant class} \\ 1, & \text{if } c \text{ is a mixed class.} \end{cases} \tag{4}$$

The type of a given class is determined based on ART2 classification of the training data set.

The structure of the ART2LDA classifier is shown in Fig. 2. The ART2 classifies the input sample x into either a malignant or a mixed class. Depending on the class type the function $g(\cdot)$ determines whether the LDA classifier will be used. If x is classified into a mixed class, the final classification will be obtained based on the LDA classifier. However, if x is classified by ART2 into a malignant class, then the mass will be considered malignant, without using the LDA classifier. Therefore, in the ART2LDA structure, the ART2 is used both as a classifier and a supervisor. This can be seen in (3). The first term in (3), $g(f_2(x))f_1(x)$, is the LDA classifier multiplied by the ART2 control part $g(f_2(x))$. The second term in (3), $(1 - g(f_2(x)))$, gives the classification result of the ART2 stage. If $f_2(x)$ is a malignant class, then $g(f_2(x)) = 0$, the LDA stage is eliminated, and the classifier output y_{AL} is equal to 1. On the other hand, if $f_2(x)$ is a mixed class, then $g(f_2(x)) = 1$, the ART2 term is eliminated, and the final classification is determined by the LDA classifier $(y_{AL} = f_1(x))$.

IV. Methods

A. Data Set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsies at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. The data set contained 348 mammograms with a mixture of benign ($n = 169$) and malignant ($n = 179$) masses. On each mammogram, a region of interest (ROI) containing the mass was identified by a radiologist experienced in breast imaging. The visibility of the masses was rated by the radiologist on a scale of 1 to 10, where the rating of 1 corresponds to the most visible category. The distributions of the visibility rating for both the malignant and benign masses are shown in Fig. 3. The visibility ranged from subtle to obvious for both types of masses. It can be observed that the benign masses tend to be more obvious than the malignant ones. Additionally the likelihood of malignancy for each mass was estimated based on its mammographic appearance. The radiologist rated the likelihood of malignancy on a scale of 1 to 10, where 1 indicated a mass with the most benign appearance. The distribution of the malignancy rating of the masses is shown in Fig. 4.

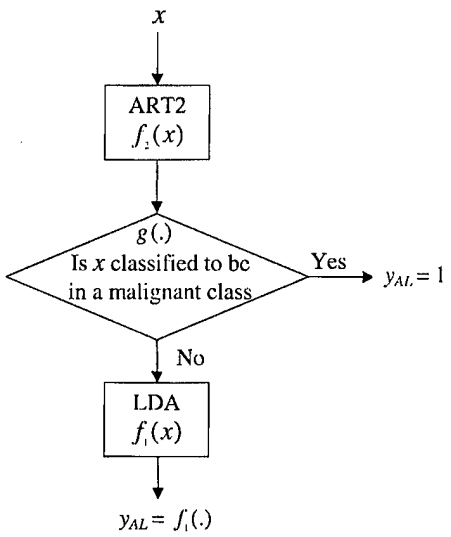


Fig. 2. Structure of the ART2LDA classifier.

The data set can be considered as representative of the patient population that is sent for biopsy under current clinical criteria. Some characteristics of many malignant and benign masses can be visually distinguished by radiologists. However, there is also a nonnegligible fraction of malignant masses that are very similar to benign masses (the low malignancy rating region in Fig. 4). The estimated likelihood of malignancy of malignant and benign masses that are sent for biopsy basically overlaps over the entire range. This is consistent with the fact that in order not to miss malignant masses radiologists must recommend biopsy for even very low suspicion lesions.

Three hundred and five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm $\times$ 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of -0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0

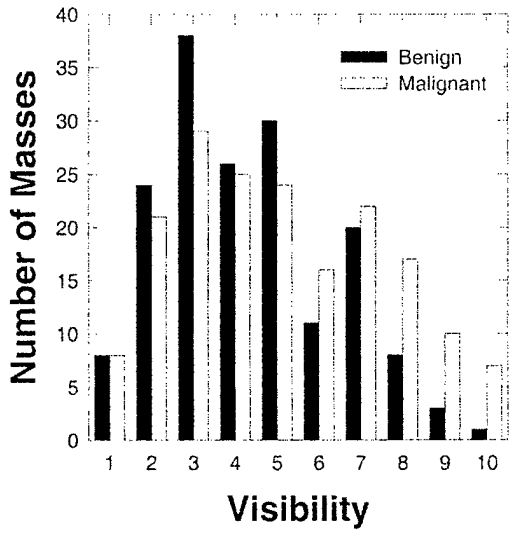


Fig. 3. The distribution of the visibility ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very obvious, 10: very subtle).

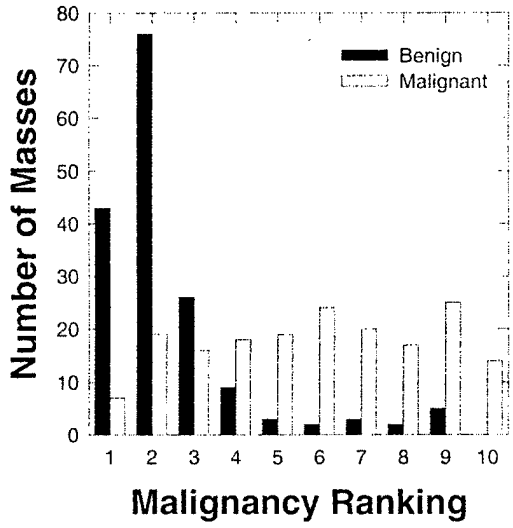


Fig. 4. The distribution of the malignancy ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very likely benign, 10: very likely malignant).

to 3.5. The remaining 43 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 50 μm $\times$ 50 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the OD within the range of 0 to 4 OD units, with a slope of -0.001 OD/pixel value. In order to process the mammograms digitized with these two different digitizers, the images digitized with LUMISCAN 85 digitizer were averaged with a 2 $\times$ 2 box filter and subsampled by a factor of two, resulting in 100 μm images.

In order to validate the prediction abilities of the classifier, the data set was partitioned randomly into training and test subsets on a 3:1 ratio, under the constraints that both the malignant and the benign samples were split with the 3:1 ratio and that the images from the same patient were grouped into the same (training or test) subset. These constraints caused the subsets to deviate from an exact 3:1 ratio. The data set was repartitioned randomly ten times. On average, 73% of the samples were grouped into the training set and 27% into the test set. The training and test results from the ten partitions were averaged to reduce their variability.

B. Feature Extraction

A rectangular ROI was defined to include the radiologist-identified mass with an additional surrounding breast tissue region of at least 40 pixels wide from any point of the mass border. A fully automated method was then used for segmentation of the mass from the breast tissue background within the ROI. The rubber band straightening transform (RBST) was previously developed [12] to map a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the border of mass appears approximately as a horizontal edge and spiculations appear approximately as vertical lines. The transformation of the radially oriented textures surrounding the mass margin to a more uniform orientation facilitates the extraction of texture features.

The texture features used in this study were calculated from spatial gray-level dependence (SGLD) matrices [10]–[12], [31], and run-length statistics (RLS) matrices [32] computed from the RBST images. The (i, j)th element of the SGLD matrix is the joint probability that gray levels i and j occur in a direction at a distance of θ pixels apart in an image. Based on our previous studies [10], a bit depth of eight was used in the SGLD matrix construction, i.e., the four least significant bits of the 12-bit pixel values were discarded. Thirteen texture measures, including correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, inertia, sum variance, difference average, difference variance, and two types of information measure of correlation were used. These measures were extracted from each SGLD matrix at ten different pixel pair distances ($d = 1, 2, 3, 4, 6, 8, 10, 12, 16$ and 20) and in four directions (0°, 45°, 90°, and 135°). Therefore, a total of 520 SGLD features were calculated for each image. The definitions of the texture measures are given in the literature [10]–[12], [31]. These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. A gray level run is a set of consecutive, collinear pixels in a given direction which have the same gray level value. The run length is the number of pixels in a run [32]. The RLS matrix describes the run length statistics for each gray level in the image. The (i, j)th element of the RLS matrix is the number of times that the gray level i in the image possesses a run length of j in a given direction. In our previous study, it was found experimentally that a bit depth of five in the RLS matrix computation could provide good texture characteristics [12].

Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity,

and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta = 0°$ and $\theta = 90°$. Therefore, a total of 20 RLS features were calculated for each ROI. The formal definition of the RLS feature measures can be found in [32].

A total of 540 features (520 SGLD and 20 RLS) were therefore extracted from each ROI.

C. Feature Selection

In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with stepwise linear discriminant analysis [33] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda (the ratio of within-group sum of squares to the total sum of squares [34]) was used as a selection criterion. The optimization procedure used a threshold F_{in} for feature entry and a threshold F_{out} for feature removal. On a feature entry step, the features not yet selected are entered into the selected feature pool one at a time, the significance of the change in the Wilks' lambda caused by this feature is estimated based on F statistics. The feature with the highest significance is entered into the feature pool if its significance is higher than F_{in}. On a feature removal step, the features which have already been selected are analyzed one at a time from the selected feature pool and the significance of the change in the Wilks' lambda is estimated. The feature with the least significance is removed from the selected feature pool if the significance is less than F_{out}. Since the appropriate values of F_{in} and F_{out} are not known *a priori*, we examined a range of F_{in} and F_{out} values and chose the appropriate thresholds in such a way that a minimum number of features were selected to achieve a high accuracy of classification by LDA for the training sets. More details about the stepwise linear discriminant analysis and its application to CAD can be found in [10]–[12].

D. Performance Analysis

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology [35]. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program [36], which fit a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. For the ART2LDA classifier, the discriminant scores of all case samples classified in the two stages are combined. All masses classified into the malignant group by the ART2 stage were assigned a constant positive discriminant score higher than or equal to the most malignant discriminant score obtained from the LDA stage .

The performance of ART2LDA was also assessed by estimation of the partial area index ($A_z^{(0.9)}$) and compared with the corresponding performance index of the LDA and BPN classifiers. The partial area index ($A_z^{(0.9)}$) is defined as the area that lies under the ROC curve but above a sensitivity threshold of 0.9 ($\text{TPF}_0 = 0.9$) normalized to the total area above TPF_0, (1-TPF_0). The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high-sensitivity (low false negative) region which is most important for clinical cancer detection task. In addition, the performance of the LDA stage of the ART2LDA classifier was evaluated by the estimation of the area under the ROC curve, denoted as A_z (LDA), for the case samples passed onto the LDA classifier.

TABLE I
NUMBER OF SELECTED FEATURES FOR THE TEN DATA GROUPS WITH THE CORRESPONDING F_{IN} AND F_{OUT} PARAMETERS

Data Group No.	Number of selected features	F_{in}	F_{out}
1	12	1.8	1.6
2	15	2.4	2.2
3	13	2.4	2.2
4	18	2.4	2.2
5	14	2.4	2.2
6	14	2.1	1.8
7	13	2.4	2.2
8	18	1.8	1.6
9	14	2.4	2.2
10	14	2.4	2.2

V. RESULTS

In this section the ART2LDA classification results for malignant and benign masses will be presented and compared with those of the LDA or BPN classifiers. The important point in this study is the fact that the test subset is truly independent of the training subset. Only the training subset is used for feature selection and classifier training, and only the test subset is used for classifier validation. In order to validate the prediction abilities of the classifier, ten different partitions of the training and test sets were used. A different ART2LDA classifier was trained using each training set and the corresponding set of selected features. The classification result was estimated as the average performance for the ten partitions.

For a given partition of training and test sets, feature selection was performed based on the training set alone. The feature selection results for the ten different training groups are shown in Table I. The average number of selected features was 14. An average of two RLS features and twelve SGLD features were selected for each of the training sets which represented 10% of all RLS features and 2.3% of all SGLD features, respectively. Both types of features (RLS and SGLD) are necessary in order to obtain good classification. The most often selected RLS features for the ten training sets were: horizontal short run emphasis (four times), horizontal long run emphasis (six times), vertical run length nonuniformity (three times), horizontal run length nonuniformity (three times). The most often selected SGLD texture measures for the ten training sets were: inverse difference moment (eight times), information measure of correlations one and two (19 times), difference average (nine times), and correlation (ten times). For a given texture measure, features at different angles or distances may be selected, but these features are usually highly correlated so

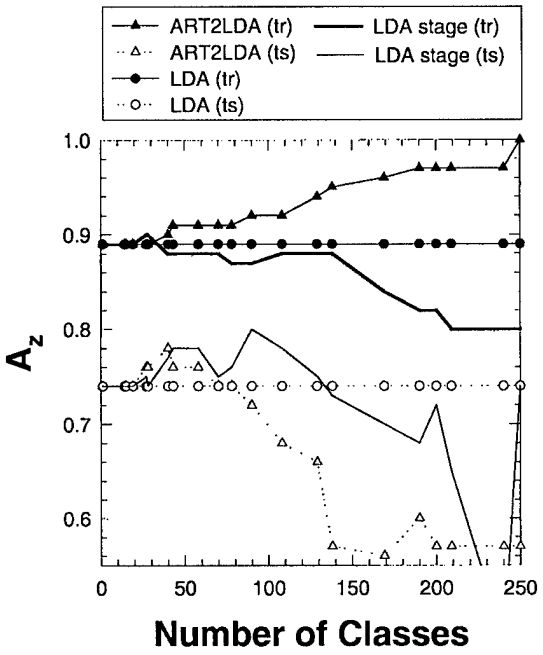


Fig. 5. ART2LDA and LDA classification results for training and test sets from data group three as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

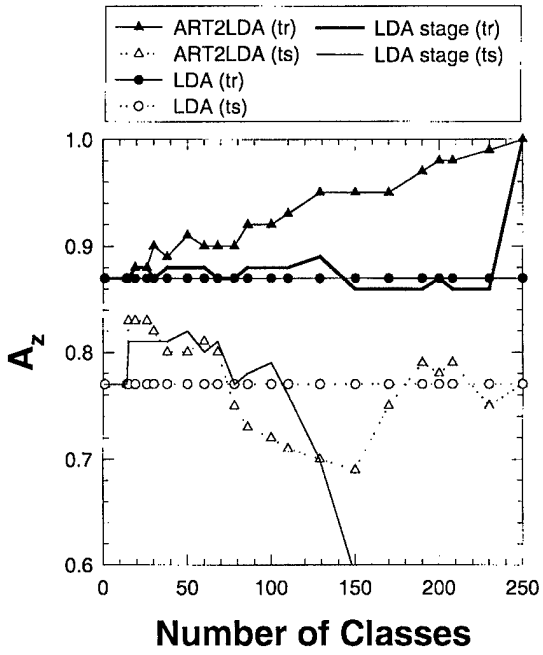


Fig. 6. ART2LDA and LDA classification results for training and test sets from data group one as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

that they can be considered to be similar and counted together as described above.

A. ART2LDA Classification Results

For the ART2LDA classifier, the number of selected features determines the dimensionality of the input vector of the ART2 classifier and the dimensionality of the LDA classifier. By applying different values for the vigilance parameter, ART2 classifiers with different number of classes were obtained. In this study, the vigilance parameter p_{vig} was varied from 0.9 to 0.99, resulting in a range of 10 to 240 classes. The overall performance of the ART2LDA classifier was evaluated for different numbers of ART2 classes because different subset of the samples were separated and classified by ART2 when p_{vig} was varied. In Fig. 5, the classification results for the ART2LDA are compared to the results from LDA alone for the training and test set partition three. The classification accuracy, A_z, was plotted as a function of the number of ART2 classes. For this training and test set partition, when the number of classes was between 20 and 60, the ART2LDA classifier improved the classification accuracy for the test set in comparison to LDA. As the number of classes increased to greater than 60, the A_z value increased for the training data set, but decreased for the test data set and was lower than that of the LDA alone. The two solid lines in Fig. 5 show the A_z values for the LDA stage in the ART2LDA classifier for both the training and test sets. It can be observed that the test A_z for the LDA stage is higher than the A_z for the LDA classifier alone, but not as high as A_z obtained by ART2LDA when the number of classes is small.

In Fig. 6 the classification results of LDA and ART2LDA for the partition one training and test sets are shown. In this case it appeared that in the test set there were two large malignant outliers which degraded the LDA performance. Only 15 classes at the ART2 stage in the ART2LDA was enough to cluster the outliers into a separate malignant class and to improve the performance of the LDA stage and the overall result. The rest of the outliers required more ART2 classes before they were clustered into separate classes and correctly classified as malignant. This is the reason for the similar behavior of the classifiers for partitions three and one in the range of 40 to 70 classes as seen in Figs. 5 and 6. When the number of classes was less than 70, the test A_z for the LDA stage (A_z(LDA)) was higher than the LDA alone, but not as high as the A_z for ART2LDA with less than 30 classes (Fig. 6). The best A_z values for the test data sets of the ten training and test partitions are presented in Table II and Fig. 7. The ART2LDA classifier achieved higher A_z values than the LDA alone in nine of the ten partitions. The average A_z is 0.81 for ART2LDA and 0.78 for LDA alone. The standard deviations of the A_z values for the ten groups range from 0.03 to 0.05 for the ART2LDA classifier and from 0.04 to 0.05 for the LDA classifier.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve $A_z^{(0.9)}$ at a TPF higher than 0.9. The results are presented in Table III and Fig. 7. In the lower part of Fig. 7, the $A_z^{(0.9)}$ values of the test set for the corresponding ten partitions of training and test sets are presented. The average test $A_z^{(0.9)}$ value is 0.34 for the ART2LDA and 0.27 for LDA. For nine of the ten partitions, the $A_z^{(0.9)}$ value was improved at the high-sensitivity operating region (TPF $>$ 0.9) of the ROC curve.

The classifier performance was also evaluated when the ART2LDA classifiers were designed using a fixed number

TABLE II
CLASSIFIERS PERFORMANCE FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE TOTAL AREA UNDER ROC CURVE

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.77	0.83	0.85	0.80
2	0.78	0.80	0.82	0.77
3	0.74	0.78	0.77	0.78
4	0.77	0.77	0.75	0.77
5	0.77	0.78	0.76	0.77
6	0.80	0.83	0.82	0.81
7	0.80	0.81	0.82	0.77
8	0.77	0.80	0.74	0.75
9	0.77	0.80	0.81	0.80
10	0.86	0.89	0.84	0.89
Mean	0.78	0.81	0.80	0.79

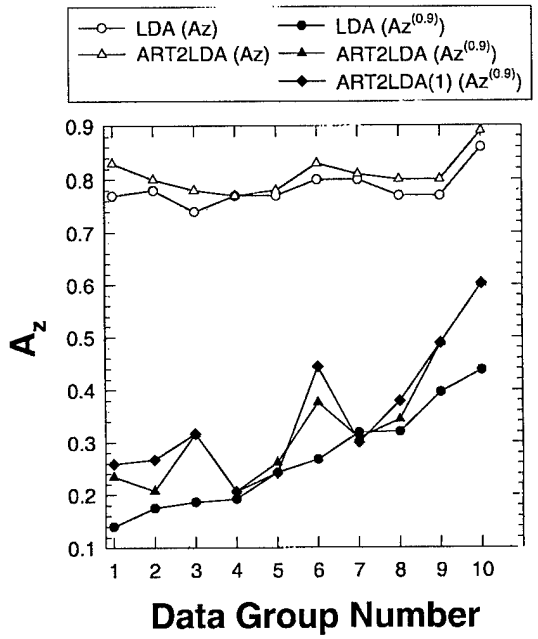


Fig. 7. Average A_z classification results for the 10 test sets. The top graphs represent the ART2LDA and LDA A_z values for the total area under the ROC curve. The bottom graphs represent the ART2LDA, ART2LDA(1) and LDA A_z values for the partial area of the ROC curve above the true positive fraction of 0.9.

TABLE III
CLASSIFIERS RESULTS FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE PARTIAL AREA OF THE ROC CURVE ABOVE THE TRUE POSITIVE FRACTION OF 0.9 ($A_z^{(0.9)}$)

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.14	0.23	0.31	0.26
2	0.17	0.21	0.28	0.27
3	0.19	0.32	0.27	0.32
4	0.19	0.21	0.19	0.21
5	0.24	0.26	0.32	0.24
6	0.27	0.38	0.27	0.44
7	0.32	0.31	0.38	0.30
8	0.32	0.34	0.25	0.38
9	0.40	0.49	0.40	0.49
10	0.44	0.60	0.38	0.60
Mean	0.27	0.34	0.31	0.35

of ART2 classes. The A_z, and $A_z^{(0.9)}$ results, averaged over the ten test partitions, are presented in Table IV. The average A_z with the ART2LDA classifier, compared to that of LDA alone, was again improved between 15 and 40 classes. The maximum average A_z of 0.80 was achieved between 20 and 40 classes. The average $A_z^{(0.9)}$ results are improved for all

TABLE IV
AVERAGE A_z AND AVERAGE $A_z^{(0.9)}$ CLASSIFICATION RESULTS FOR THE TEN TEST SETS. CLASSIFIERS WERE DESIGNED USING A FIXED NUMBER OF ART2 CLASSES

	LDA	ART2LDA					
No. of classes		15	20	30	40	50	60
A_z	0.78	0.80	0.80	0.80	0.80	0.78	0.77
$A_z^{(0.9)}$	0.27	0.30	0.31	0.33	0.33	0.31	0.31

ART2LDA classifiers presented in Table IV. The maximum average $A_z^{(0.9)}$ value is 0.33 and it remains constant between 30 and 40 classes.

An alternative way to evaluate the performance of a classifier is its classification accuracy when a decision threshold for malignancy is selected based on the training set. For instance, a decision threshold may be selected such that all positive samples from the training set are classified correctly i.e., at a sensitivity of 100%. The ART2LDA with this decision threshold is referred to as ART2LDA(1). For a given training and test partitioning, ART2LDA classifiers with different number of classes in the ART2 stage were obtained (Figs. 5 and 6). For each of these models the decision threshold for a sensitivity of 100% was selected from the training set and the corresponding ART2LDA(1) classifier was obtained. Then the ART2LDA(1) classifier (with a specific number of classes in the ART2 stage) that correctly classified the maximum number of malignant masses in the test set is selected. By using all samples of the test set, the A_z value is calculated for the corresponding ART2LDA model. The A_z values for the ART2LDA(1) classifiers for the test sets of the ten data partitionings are shown in Tables II and III. For five of the partitions the overall A_z value for ART2LDA(1) is higher than that of LDA alone (Table II). The average A_z value was 0.79. The partial areas above the TP fraction of 0.9, $A_z^{(0.9)}$, for the ten test data sets obtained by the ART2LDA(1) classifier are also shown in Fig. 7. The ART2LDA(1) achieved the highest average $A_z^{(0.9)}$ value of 0.35 compared to ART2LDA and LDA (Table III).

B. BPN Classification Results

A multilayer perceptron back-propagation neural network with a single hidden layer and a single output node was used for comparison with the ART2LDA classifier. The number of selected features determined the number of input nodes to the BPN. The same ten training/test set partitions (as in the case of ART2LDA) were used for the training and validation of the BPN classifiers. BPN's with their number of hidden nodes ranging from two to ten were evaluated to obtain the best architecture. Back-propagation training was used. Each of the BPN's was trained for up to 18 000 training epochs. At every 1000 epochs the neural network weights were saved and the classification result for the corresponding test set was evaluated. This design procedure was repeated for each of the ten training/test groups. For each group, the best test result among all the BPN architectures (different number of hidden nodes) and all the training epochs examined was selected. The average test A_z over the ten groups for the BPN was 0.80, compared to 0.81 for ART2LDA (Table II). The standard deviations of the A_z values for the ten groups range from 0.04 to 0.05 for the BPN. The average partial $A_z^{(0.9)}$ for the BPN

was 0.31, compared to 0.34 for ART2LDA (Table III). The A_z and $A_z^{(0.9)}$ of the ART2LDA classifier were higher than those of the BPN in six of the ten training/test groups.

VI. Discussion

In the present study, a new classifier (ART2LDA) was designed and applied to the classification of malignant and benign masses. The results indicated that the ART2LDA classifier had better generalizability than an LDA classifier alone. The ART2 classifier grouped the case samples that were different from the main population into separate classes. The minimum number of classes needed to start the clustering of outliers into separate classes depended on how different the outliers were from the rest of the sample population. For the ten different partitions of training and test sets used in this study, the minimum number varied between 13 and 15 classes. When the number of ART2 classes was less than this minimum number of classes, the ART2 classifier generated only mixed malignant-benign classes and all samples were transferred to the LDA stage. In that case, the ART2LDA was equivalent to the LDA classifier alone. When a higher number of classes were generated, an increased number of cases that might be considered outliers of the general data population was removed (clustered in separate classes). For the ten training sets used in this study, the malignant outliers were gradually removed when the number of classes increased. The training accuracy increased when the number of classes increased and A_z could reach the value of 1.0. However, a large number of ART2 classes led to overfitting the training sample set and poor generalization in the test set. The classification accuracy of ART2 for the test set tended to decrease when the number of classes was greater than about 70. The large number of classes also led to a reduction in the generalizability of the second-stage LDA; the training of LDA with a small number of samples would again result in overfitting the training set, and poor generalizability in the test set. This effect was observed when more than 60 or 70 classes were generated by ART2 (see Figs. 5 and 6).

The classification accuracy of ART2LDA increased initially with an increased number of classes and then decreased after reaching a maximum. The correct classification of the outliers by the ART2 in combination with an improvement in the classification by the LDA resulted in the increased accuracy. When the number of ART2 classes was further increased, the effects of overfitting by the ART2 and the LDA became dominant and the prediction ability of the ART2LDA decreased. In some cases the second-stage LDA prediction was much worse than the ART2. In other cases the ART2 could not generalize well. The generation of a high number of classes is therefore impractical and unnecessary both from a computational and a methodological point of view.

For the optimal number of classes (usually less than 50 for the data sets used) the A_z value for the second-stage LDA in the ART2LDA was better than an LDA classifier alone, but it was not as good as the overall A_z from the ART2LDA. It is evident that the ART2 was a useful classifier for improvement of the second-stage classification.

When the partial area of the ROC curve above the true positive fraction (TPF) of 0.9 ($A_z^{(0.9)}$) was considered as a measure of classification accuracy, the advantage of ART2LDA over LDA alone became even more evident. By removing and correctly classifying the outliers, the accuracy of the classification was increased at the high sensitivity end of the curve.

The classifier performance was evaluated when the ART2LDA classifiers were designed using a fixed number of ART2 classes. The results showed improved performance of the ART2LDA in a range between 20 and 40 ART2 classes. Both the average A_z and the average $A_z^{(0.9)}$ reached a maximum within this region, and the maximum average A_z and the average $A_z^{(0.9)}$ values remained unchanged between 30 and 40 classes. These results indicated that the performance of a hybrid ART2LDA classifer was robust and stable and could be potentially useful in real clinical applications.

We have performed statistical tests with the CLABROC program to estimate the significance in the differences between the A_z values from the ART2LDA, the LDA alone, and the BPN, as well as in the differences in the partial $A_z^{(0.9)}$ from the three classifiers. The statistical tests were performed for each individual data set partition because the correlation among the data sets from the different partitions precludes the use of student's paired t test with the ten partitions. We found that the differences in both cases did not reach statistical significance because of the small number of test samples and thus the large standard deviation in the A_z values. However, the consistent improvements in A_z and $A_z^{(0.9)}$ by the ART2LDA (9 out of 10 data set partitions in both cases for LDA and six out of ten data set partitions in both cases for BPN) suggest that the improvement was not by chance alone, and that the accuracy of a classification task could be improved by the use of an ART2 network. In addition, one advantage of the ART2LDA is that the training process is more efficient than that of the BPN, especially when there is a subset of outlying samples. In such a case, the BPN will require a large number of training epochs to minimize the error function.

ART2LDA can be trained to classify the sample cases into more than two classes, such as a class of normal tissue regions in addition to malignant and benign masses. There will be an increase in the complexity of training and a larger training sample size will be desired, but these requirements will be comparable for the different classifiers. In a clinical situation, if the classification task is performed on all computer-detected lesions, the classifier has to distinguish the falsely detected normal tissue from malignant or benign lesions. However, it may be noted that a classifier that can distinguish only malignant and benign masses is applicable to the scenario that the radiologist identifies a suspicious lesion on the mammogram and would like to have a second opinion about its likelihood of malignancy before making a diagnostic decision. Therefore, the development of a classifier that can differentiate malignant and benign masses is the research of interest for many investigators.

Similarly, ART2 can be trained to discover and remove a pure benign mass class. The approach will be similar to the task of classifying and removing the pure malignant classes,

as described in this study. However, our approach of removing the malignant classes will reduce the chance of misclassification of malignant masses. In breast cancer detection, the cost of false-negative (missed cancer) is very high. Therefore, our goal in classifier design is to be conservative. By removing the malignant classes in the first stage, any misclassification to these classes will be regarded as malignant. The remaining classes will be classified again with the second-stage classifier so malignant masses will be less likely to be missed.

The problem of classification of malignant and benign masses has been studied by many investigators. Rangayyan *et al.* [15] used Mahalanobis distance classifer (a modification of an LDA classifier) and the leave-one-out method to evaluate the classification of 54 masses. Fogel *et al.* [16] compared LDA and BPN classifiers using the leave-one-out method and 139 masses (malignant and benign classification). Highnam *et al.* [17] used a morphological feature called a halo to classify 40 masses as malignant and benign. Huo *et al.* [22] employed BPN and a rule-based classifier to classify 95 masses using the leave-one-out evaluation method. Sahiner *et al.* [12] used an LDA classifier and the leave-one-out method to classify 168 masses. An important difference between the classifier designed in this study and the previous studies in the CAD field is the method of feature selection. In the above mentioned studies [12], [15]–[17], [22] and several other published studies [18]–[21] the features were selected from the entire data set first, and then the data set was partitioned into training and test sets. This meant that at the feature selection stage of the classifier design, the entire data set was used as a training set. Depending on the distribution of the features and the total number of samples used, the test results in these studies might be optimistically biased [37]. In our current study, the entire data set was initially partitioned into training and test sets and then feature selection was performed only on the training set. This method will result in a pessimistic estimate of the classifier performance when the training set is small [37]. However, it will provide a more conservative but realistic estimation of the classifier performance in the general patient population. We can expect that the performance would be improved if the classifier in this study were designed using a large data set. Since our main purpose in this study was to compare the ART2LDA classifier with the commonly used LDA and BPN, we did not attempt to quantify how pessimistic our results were in this study.

The most important contribution of this paper is to introduce a new approach that utilizes a two-stage unsupervised–supervised hybrid classifier. We believe that the hybrid approach will improve classification when the sample distribution contains subpopulations that may be difficult for a single classifier to classify. It will be useful for similar classification tasks although different classifiers may be used in each stage of the hybrid structure.

VII. Conclusion

A new classifier combining an unsupervised ART2 and a supervised LDA has been designed and applied to the classification of malignant and benign masses. A data set consisting of 348 films (179 malignant and 169 benign) was randomly partitioned into training and test subsets. Ten different random partitions were generated. For each training set, texture features were extracted and feature selection was performed. An average of features were selected for each group. A hybrid ART2LDA classifier, an LDA, and a BPN were trained by using each of the ten training sets. The A_z value under the ROC curve for the test sets, averaged over the ten partitions, was higher for ART2LDA ($A_z = 0.81$) compared to those of the LDA alone ($A_z = 0.78$) and of the BPN ($A_z = 0.80$). A greater improvement was obtained when the partial ROC area above a true-positive fraction of 0.9 was considered. The average partial A_z for ART2LDA was 0.34, as compared to 0.27 for LDA and 0.31 for BPN. Additionally, for the ART2LDA classifiers that correctly classified the maximum number of malignant masses in the test sets with decision threshold defined with the training set, the average partial A_z was 0.35. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

Acknowledgment

The authors would like to thank Prof. S. Grosberg and Dr. G. Carpenter for providing them with valuable information as well as for the useful discussions. Additionally the authors would like to thank C. E. Metz, Ph.D., for providing the LABROC1 and CLABROC programs.

References

[1] H. C. Zuckerman, "The role of mammography in the diagnosis of breast canser," in *Breast Canser, Diagnosis and Treatment*, I. M. Ariel and J. B. Cleary, Eds. New York: McGraw-Hill, 1987, pp. 152–172.
[2] D. B. Kopans, "The positive predictive value of mammography," *Amer. J. Roentgenol.*, vol. 158, pp. 521–526, 1992.
[3] D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," *Curr. Opin. Radiol.*, vol. 4, pp. 123–129, 1992.
[4] M. Moskowitz, "Impact of a priory medical detection on screening for breast cancer," *Radiology*, vol. 184, pp. 619–622, 1989.
[5] P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner, 1975.
[6] R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[7] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1974.
[8] D. Rumelhart, G. E. Hinton, and R. J. Williams, in D. E. Rumelhart, Ed., *Parallel and Distributed Processing*. Cambridge, MA: MIT Press, 1986, vol. 1, p. 318.
[9] J. Herz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1991.
[10] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminat analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857–876, 1995.
[11] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Med. Phys.*, vol. 22, pp. 1501–1513, 1995.
[12] B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mamograms: The rubber band sraightening transform and texture analysis," *Med. Phys.*, vol. 25, no. 4, pp. 516–526, Apr. 1998.
[13] B. Sahiner, H. P. Chan, D. Wei, N. Petick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Med. Phys.*, vol. 23, no. 10, pp. 1671–1683, Oct. 1996.
[14] H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant

and benign microcalsifications on mammograms: Texture analysis using an artificial neural network," *Phys. Med. Biol.*, vol. 42, pp. 549–567, 1997.

[15] R. M. Rangayyan, N. M. El-Farmawy, J. E. Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Trans. Med. Imag.*, vol. 16, pp. 799–810, Dec. 1997.

[16] D. B. Fogel, E. C. Wasson, E. M. Boughton, V. W. Porto, and P. J. "Angeline, linear and neural model for classifying breast masses," *IEEE Trans. Med. Imag.*, vol. 17, pp. 485–488, June 1998.

[17] R. P. Highnam, J. M. Brady, and B. J.Shepstone, "A quantitative feature to aid diagnosis in mammography," in *Proc. Digital Mammography'96*, pp. 201–206.

[18] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, 1993.

[19] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvements in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," *Med. Phys.*, vol. 19, pp. 1475–1481, 1992.

[20] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," *IEEE Trans. Med. Imag.*, vol. 12, pp. 664–669, Dec. 1993.

[21] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," *IEEE Trans. Med. Imag.*, vol. 14, pp. 537–547, Sept. 1995.

[22] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[23] M. Jordan, and R. A. Jacobs, "Hierarchical mixture of experts and EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[24] L. Hadjiiski and P. Hopke, "Design of large scale models based on multiple neural network approach," *Intelligent Engineering Systems Through Artificial Neural Networks*. ASME, 1997, vol. 7, pp. 61–66.

[25] S. Grossberg, "Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors," *Biolog. Cybern.*, vol. 23, no. 3, pp. 121–134, 1976.

[26] G. A. Carpenter and S. Grossberg, "ART 2: Self-organization of stable category recognition codes for analog input patterns," *Appl. Opt.*, vol. 26, no. 23, 1, pp. 4919–4930, Dec. 1987.

[27] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition," *Neural Networks*, vol. 4, no. 4, pp. 493–504, 1991.

[28] G. A. Carpenter and S. Grossberg, "Integrating symbolic and neural processing in a self-organizing architeture for pattern recognition and prediction," in *Artificial Intelligence and Neural Networks: Steps toward Principled Integration*. New York: Academic, 1994.

[29] G. A. Carpenter and N. Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," *Neural Networks*, vol. 11, no. 2, pp. 323–336, Mar. 1998.

[30] Y. Xie, P. K. Hopke, and D. Wienke, "Airborne particle classification with a combination of chemical composition and shape index utilizing an adaptive resonance artificial neural network," *Environ. Sci. Technol.*, vol. 28, no. 11, pp. 1921–1928, 1994.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, pp. 610–621, Nov. 1973.

[32] M. M. Galloway, "Texture analysis using gray level run length," *Comput. Graph. Image Processing*, vol. 4, pp. 172–179, 1975.

[33] M. J. Norusis, *SPSS Professional Statistics 6.1*. Chicago, IL: SPSS, 1993.

[34] M. M. Tatsuoka, "Multivariate Analysis," *Techniques for Educational and Psychological Research*. New York: Macmillan, 1988.

[35] C. E. Metz, "ROC methodology in radiographic imaging," *Invest. Radiol.*, vol. 21, pp. 720–733, 1986.

[36] C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binomial ROC curve from continuously distributed test results," presented at the *1990 Annu. Meeting American Statistical Association*, Anahaim, CA, 1990.

[37] B. Sahiner, H. P. Chan, N. Petrick, R. Wagner, and L. Hadjiiski, "The effect of sample size on feature selection in computer-aided diagnosis," *Proc. SPIE*, vol. 3661, pp. 499–510, 1999.

Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size

Berkman Sahiner,[a] Heang-Ping Chan, Nicholas Petrick,
Robert F. Wagner,[b] and Lubomir Hadjiiski
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0904

(Received 1 July 1999; accepted for publication 18 April 2000)

In computer-aided diagnosis (CAD), a frequently used approach for distinguishing normal and abnormal cases is first to extract potentially useful features for the classification task. Effective features are then selected from this entire pool of available features. Finally, a classifier is designed using the selected features. In this study, we investigated the effect of finite sample size on classification accuracy when classifier design involves stepwise feature selection in linear discriminant analysis, which is the most commonly used feature selection algorithm for linear classifiers. The feature selection and the classifier coefficient estimation steps were considered to be cascading stages in the classifier design process. We compared the performance of the classifier when feature selection was performed on the design samples alone and on the entire set of available samples, which consisted of design and test samples. The area A_z under the receiver operating characteristic curve was used as our performance measure. After linear classifier coefficient estimation using the design samples, we studied the hold-out and resubstitution performance estimates. The two classes were assumed to have multidimensional Gaussian distributions, with a large number of features available for feature selection. We investigated the dependence of feature selection performance on the covariance matrices and means for the two classes, and examined the effects of sample size, number of available features, and parameters of stepwise feature selection on classifier bias. Our results indicated that the resubstitution estimate was always optimistically biased, except in cases where the parameters of stepwise feature selection were chosen such that too few features were selected by the stepwise procedure. When feature selection was performed using only the design samples, the hold-out estimate was always pessimistically biased. When feature selection was performed using the entire finite sample space, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, the number of available samples, and their statistical distribution. For our simulation conditions, these estimates were always pessimistically (conservatively) biased if the ratio of the total number of available samples per class to the number of available features was greater than five. [S0094-2405(00)01607-2]

Key words: feature selection, linear discriminant analysis, effects of finite sample size, computer-aided diagnosis

I. INTRODUCTION

Computer-aided interpretation of medical images has been the subject of numerous studies in recent years. The purpose of computer-aided diagnosis (CAD) in medical imaging is to provide a second opinion to the radiologist concerning the presence or the likelihood of malignancy of abnormalities in a given image or case. The general visual criteria that help describe the abnormality or its classification can usually be provided by the radiologist. However, in many cases, it is difficult to translate these criteria into computer algorithms that exactly match the verbal description of what the radiologist visually perceives. Therefore, a common first step in CAD is to extract a number of features, or a feature space, that is believed to have a potential for the given task. The features may or may not match to what a radiologist searches in the image for the same task. In the next step, a subset of features are selected from the entire feature space based on their individual or joint performance, and the selected set of features are used in the remaining steps of the CAD system. This approach also has the advantage that the computer may discover some features that are difficult to perceive or verbally describe by the radiologist, so that the computer may extract information that is complementary to the radiologist's perceived image features.

A common problem in CAD is the lack of a large number of image samples to design a classifier and to test its performance. Although the effect of finite sample size on classification accuracy has previously been studied, many elements of this research topic warrant further study. In order to treat specific components of this problem, previous studies have mostly ignored the feature selection component of this problem, and assumed that the features to be used in the classifier have been chosen and are fixed.[1–6] However, as described in the previous paragraph, feature selection is a necessary first step in many CAD algorithms. This paper addresses the effect of finite sample size on classification accuracy when the classifier design involves feature selection.

When only a finite number of samples are available for

classifier design and testing, two commonly used performance estimates are those provided by the resubstitution and the hold-out methods. In the hold-out method, the samples are partitioned into independent training and test samples, the classifier is designed using the training samples alone, and the accuracy of the designed classifier is measured by its performance for the test samples. In the resubstitution method, the accuracy is measured by applying the classifier to the training samples that have been used to design it. Other methods such as leave-one-out and bootstrap have also been shown to be very useful procedures for performance estimation with a finite sample size.[7] As the number of training samples increases, all of these estimates approach the true classification accuracy, which is the accuracy of a classifier designed with the full knowledge of the population distributions. When the training sample size is finite, it is known that, on average, the resubstitution estimate of classifier accuracy is optimistically biased relative to that of a classifier trained with an infinite sample. In other words, it has a higher expected value than the performance obtained with an infinite design sample set, which is the true classification accuracy. Similarly, on average, the hold-out estimate is pessimistically biased, i.e., it has a lower expected value than the true classification accuracy. When classifier design is limited by the availability of design samples, it is important to obtain a realistic estimate of the classifier performance so that classification will not be misled by an optimistic estimate such as that provided by resubstitution.

In CAD literature, different methods have been used to estimate the classifier accuracy when the classifier design involves feature selection. In a few studies, only the resubstitution estimate was provided.[8] In some studies, the researchers partitioned the samples into training and test groups at the beginning of the study, performed both feature selection and classifier parameter estimation using the training set, and provided the hold-out performance estimate.[9,10] Most studies used a mixture of the two methods. The entire set of available samples was used as the training set at the feature selection step of classifier design. Once the features have been chosen, the hold-out or leave-one-out methods were used to measure the accuracy of the classifier.[11–16] To our knowledge, it has not been reported whether this latter method provides an optimistic or pessimistic estimate of the classifier performance.

A powerful method for estimating the infinite-sample performance of a classifier using a finite number of available samples was first suggested by Fukunaga and Hayes.[17] In the Fukunaga–Hayes method, subsets of $N_1, N_2, \ldots, N_j$ design samples are drawn from the available sample set, the classifier accuracy is evaluated at these different sample sizes, and the infinite-sample performance is estimated by linear extrapolation from the j points to $N \to \infty$ or $1/N \to 0$. This method has recently been applied to performance estimation in CAD, where the area A_z under the receiver operating characteristic (ROC) curve is commonly used as the performance measure.[1–3] For various classifiers and Gaussian sample distributions, the A_z value was plotted against $1/N_i$, and it was observed that the dependence of the A_z value can be closely approximated by a linear relationship in a sample size range where higher-order terms in $1/N_i$ can be neglected.[1–3] This facilitates estimation of the infinite-sample performance from the intercept of a linear regression.

This paper describes a simulation study that investigates the effect of finite sample size on classifier accuracy when classifier design involves feature selection using stepwise linear discriminant analysis. The classification problem was defined as deciding whether a sample belongs to either one of two classes, and the two classes were assumed to have multivariate Gaussian distributions with equal covariance matrices. We chose to focus our attention on stepwise feature selection in linear discriminant analysis since this is a commonly used feature selection and classification method. The effects of different covariance matrices and means on feature selection performance were studied. We examined the effects of sample size, number of available features, and parameters of stepwise feature selection on classifier bias. The biases of the classifier performance when feature selection was performed on the entire sample space and on the design samples alone were compared. Finally, we investigated whether the methods of infinite-sample performance estimation developed previously[1–3,17] can be applied to our problem.

II. METHODS

In our approach, the problem of classifier design is analyzed in two stages. The first stage is stepwise feature selection, and the second stage is the estimation of the coefficients in the linear discriminant formulation using the selected feature subset as predictor variables.

A. Stepwise feature selection

The two-class classification defined in the last paragraph of the Introduction can be formulated as a first-order linear multiple regression problem.[18] Since most of the literature on stepwise feature selection is based on the linear regression formulation, we will use this formulation to describe stepwise feature selection in this subsection. A different statistical formulation of the problem, which coincides with the linear regression formulation if the covariance matrices of the classes are equal,[18] will be described in Sec. II A, and will be used in the remainder of the paper.

Let N denote the number of samples available to design the classifier, and let k denote the number of features. In the linear multiple regression formulation, a desired output $o(i)$ is assigned to each k-dimensional feature vector X_i such that

$$o(i)=\begin{cases} o_1 & \text{if } i \in \text{class } 1 \\ o_2 & \text{if } i \in \text{class } 2 \end{cases}. \tag{1}$$

To define the linear multiple regression problem, the desired outputs $o(i)$ are used as the dependent variable and the feature vectors X_i are used as the independent variables. The discriminant score for a feature vector X_i is the predicted value of $o(i)$, computed by the regression equation

$$h^{(k)}(X_i)=b^T X_i + b_0, \tag{2}$$

where $b^T=[b_1,b_2,\ldots,b_k]$ and b_0 are the regression coefficients. Stepwise feature selection iteratively changes the number of features k used in the classification by entering features into or removing features from the group of selected features based on a feature selection criterion using F-statistics.[19,20] We have used stepwise feature selection for classifier design in many of our CAD applications.[11,21–23] In this study, Wilks' lambda, which is defined as the ratio of within-group sum of squares to the total sum of squares of the discriminant scores, was used as the feature selection criterion. Let $m_1^{(k)}$ and $m_2^{(k)}$ denote the means of the discriminant scores for classes 1 and 2, respectively, and let $m^{(k)}$ denote the mean of the discriminant scores computed over both classes. Wilks' lambda λ_k is defined as[19]

$$\lambda_k=\frac{\Sigma_{i\in\text{class }1}(h^{(k)}(X_i)-m_1^{(k)})^2+\Sigma_{i\in\text{class }2}(h^{(k)}(X_i)-m_2^{(k)})^2}{\Sigma_{i=1}^{N}(h^{(k)}(X_i)-m^{(k)})^2}. \quad (3)$$

A smaller value for Wilks' lambda means that the spread within each class is small compared with the spread of the entire sample, which means the separation of the two classes is relatively large and that better classification is possible. Entering a new feature into regression will always decrease Wilks' lambda, unless the feature is completely useless for classifying the available samples. The problem is to decide whether the decrease in Wilks' lambda justifies entering the feature into regression. In stepwise feature selection an *F-to-enter* value—for making the decision whether a feature should be entered when k features are already used—is defined as[24]

$$F=(N-k-2)\left(\frac{\lambda_k}{\lambda_{k+1}}-1\right), \quad (4)$$

where λ_k is Wilks' lambda before entering the feature, and λ_{k+1} is Wilks' lambda after entering the feature. An *F-to-remove* value is similarly defined to decide whether a feature already in the regression should be removed. At the feature entry step of the stepwise algorithm, the feature with the largest *F-to-enter* value is entered into the selected feature pool if this maximum value is larger than a threshold F_{in}. At the feature removal step, the feature with the smallest *F-to-remove* value is removed from the selected feature pool if this minimum value is smaller than a threshold F_{out}. The algorithm terminates when no more features can satisfy the criteria for either entry or removal. The number of selected features increases, in general, when F_{in} and F_{out} are reduced.

In order to avoid numerical instabilities in the solution of linear systems of equations, a tolerance term is also employed in the stepwise procedure to exclude highly correlated features. If the correlation between a new feature and the already selected features is larger than a tolerance threshold, then the feature will not be entered into the selected feature pool even if it satisfies the feature entry criterion described in the previous paragraph.

Since the optimal values of F_{in} and F_{out} for a given classification task are not known *a priori*, these thresholds have to be varied over a range in order to find the "best" combinations of features in a practical application. In this simulation study, we limit our selection of F_{out} to $F_{\text{out}}=F_{\text{in}}-1$, so that we do not search through all combinations of F values. This constraint should not limit our ability to demonstrate the effect of finite sample size on feature selection and classifier performance, because we were still able to vary the number of selected features over a wide range, as will be shown in Figs. 6 and 12 below.

B. Estimation of linear discriminant coefficients

As a by-product of the stepwise feature selection procedure used in our study, the coefficients of a linear discriminant classifier that classifies the design samples using the selected features as predictor variables are also computed. However, in this study, the design samples of the stepwise feature selection may be different from those used for coefficient estimation in the linear classifier. Therefore, we implemented the stepwise feature selection and discriminant coefficient estimation components of our classification scheme separately.

Let Σ_1 and Σ_2 denote the k-*by*-k covariance matrices of samples belonging to class 1 and class 2, and let $\mu_1=(\mu_1(1),\mu_1(2),\ldots,\mu_1(k))$, $\mu_2=(\mu_2(1),\mu_2(2),\ldots,\mu_2(k))$ denote their mean vectors. For an input vector X, the linear discriminant classifier output is defined as

$$h(X)=(\mu_2-\mu_1)^T\Sigma^{-1}X+\tfrac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1-\mu_2^T\Sigma^{-1}\mu_2), \quad (5)$$

where $\Sigma=(\Sigma_1+\Sigma_2)/2$. Because of the assumption in this study that the two covariance matrices are equal, Σ reduces to $\Sigma=\Sigma_1=\Sigma_2$. Therefore, we will be concerned with only the form of Σ in the following discussions. The linear discriminant classifier is the optimal classifier when the two classes have a multivariate Gaussian distribution with equal covariance matrices.

For the class separation measures considered in this paper (refer to Sec. II C), the constant term $(\mu_1^T\Sigma^{-1}\mu_1-\mu_2^T\Sigma^{-1}\mu_2)/2$ in Eq. (1) is irrelevant. Therefore, the classifier design can be viewed as the estimation of k parameters of the vector $b=(\mu_2-\mu_1)^T\Sigma^{-1}$ using the design samples.

When a finite number of design samples are available, the means and covariances are estimated as the sample means

and the sample covariances from the design samples. The substitution of the true means and covariances in Eq. (1) by their estimates causes a bias in the performance measure of the classifier. In particular, if the designed classifier is used for the classification of design samples, then the performance is optimistically biased. On the other hand, if the classifier is used for classifying test samples that are independent from the design samples, then the performance is pessimistically biased.

C. Measures of class separation

The traditional assessment methodology in medical imaging is receiver operating characteristic (ROC) analysis, which was first developed in the context of signal detection theory.[25–27] In this study, the output score of the classifier was used as the decision variable in ROC analysis, and the area A_z under the ROC curve was used as the principal measure of class separation. Excellent reviews of ROC methods applied to medical imaging can be found in the literature.[28–30]

1. Infinite sample size

When an infinite sample size is available, the class means and covariance matrices can be estimated without bias. In this case, we use the squared Mahalanobis distance $\Delta(\infty)$, or the area $A_z(\infty)$ under the ROC curve as the measures of class separation, as explained below. The infinity sign in parentheses denotes that the distance is computed using the true means and covariance matrices, or, equivalently, using an infinite number of random samples from the population.

Assume that two classes with multivariate Gaussian distributions and equal covariance matrices have been classified using Eq. (1). Since Eq. (1) is a linear function of the feature vector X, the distribution of the classifier outputs for class 1 and class 2 will be Gaussian. Let m_1 and m_2 denote the means of the classifier output for the case of the normal class, and for the case of the abnormal class, respectively, and let s_1^2 and s_2^2 denote the variances. With the squared Mahalanobis distance $\Delta(\infty)$ defined as

$$\Delta(\infty)=(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1), \tag{6}$$

it can be shown that

$$m_2-m_1=s_1^2=s_2^2=\Delta(\infty). \tag{7}$$

The quantity $\Delta(\infty)$ is referred to as the squared Mahalanobis distance between the two classes. It is the square of the Euclidean distance between the two classes, normalized to the common covariance matrix.

In particular, if Σ is a k-by-k diagonal matrix with $\Sigma_{i,i}=\sigma^2(i)$, then

$$\Delta(\infty)=\sum_{i=1}^{k}\delta(i), \tag{8}$$

where

$$\delta(i)=[\mu_2(i)-\mu_1(i)]^2/\sigma^2(i) \tag{9}$$

is the squared signal-to-noise ratio of the distributions of the two classes for the ith feature.

Using Eq. (3), and the normality of the classifier outputs, it can be shown that[31]

$$A_z(\infty)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\sqrt{\Delta/2}}e^{-t^2/2}dt. \tag{10}$$

2. Finite sample size

When a finite sample size is available, the means and covariances of the two class distributions are estimated as the sample means and the sample covariances using the design samples. The output score of the linear discriminant classifier for a test sample is computed using Eq. (1). The accuracy of the classifier in discriminating the samples from the two classes is measured by ROC methodology. The discriminant score is used as the decision variable in the LABROC program,[32] which provides the ROC curve based on maximum likelihood estimation.[33]

D. Simulation conditions

In our simulation study, we assumed that the two classes follow multivariate Gaussian distributions with equal covariance matrices and different means. This assumption is an idealization of the real class distributions that one may observe in a practical classification problem. It restricts the number of parameters in our simulations to a manageable range, while permitting us to approximate a range of situations that may be encountered in CAD.

We generated a set of N_s samples from each class distribution using a random number generator. The sample space was randomly partitioned into N_t training samples and N_s-N_t test samples per class. For a given sample space, we used several different values for N_t in order to study the effect of the design sample size on classification accuracy. For a given N_t, the sample space was independently partitioned 20 times into N_t training samples and N_s-N_t test samples per class, and the classification accuracy A_z obtained from these 20 partitions was averaged in order to reduce the variance of the classification accuracy estimate. The procedure described above was referred to as one experiment. For each class distribution described in Cases 1, 2, and 3 below, 50 statistically independent experiments were performed, and the results were averaged.

Two methods for feature selection were considered. In the first method, the entire sample space with N_s samples per class was used for feature selection. In other words, the entire sample space was treated as a training set at the feature selection step of classifier design. After feature selection, the training-test partitioning was used to evaluate the resubstitution and hold-out performances of the coefficient estimation step of classifier design. In the second method, both feature selection and coefficient estimation were performed using only the training set with N_t samples per class.

Case 1: Identity covariance matrix

In the first simulation condition, a hypothetical feature space was constructed such that the covariance matrices of

the two classes $\Sigma_1=\Sigma_2=\Sigma$ was the identity matrix, and the mean difference $\Delta\mu$ between the two classes for feature i was

$$\Delta\mu(i)=\mu_2(i)-\mu_1(i)=\alpha\beta^i, \quad i=1,...,M \quad \text{and} \quad \beta<1, \tag{11}$$

where M refers to the number of available features for feature selection. Note that k, previously defined in Sec. II B, refers to the number of features selected for classifier parameter estimation; therefore, in general, $M \geq k$. For a given data set, the number of available features M is fixed, whereas the number of selected features k depends on the F_{in} and F_{out} parameters of the stepwise selection algorithm. Since β is chosen to be less than 1, the ability for separation of the two classes by feature no. i decreased as i increased, as evidenced by $\delta(i)=(\alpha\beta^i)^2$ [see Eq. (5)]. The squared Mahalanobis distance $\Delta(\infty)$ was computed as

$$\Delta(\infty)=\frac{\alpha^2\beta^2}{1-\beta^2}(1-\beta^{2M})$$

since $\sigma(i)=1$ for all i's.

In our simulation, we chose $\beta=0.9$, and chose α such that $\Delta(\infty)=3.0$, or $A_z(\infty)=0.89$. The value of $A_z(\infty)$ versus k is plotted in Fig. 1, when features 1 through k were included in the linear discriminant. It is seen that for $k>25$, the contribution of an additional feature to the classification accuracy was very close to zero. With this simulation condition, we studied the classification accuracy for three different numbers of available features, namely, $M=50$, $M=100$, and $M=200$.

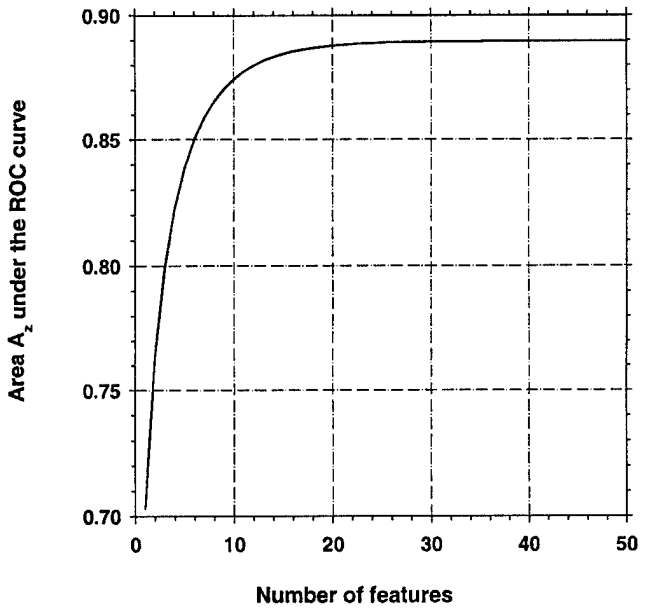


FIG. 1. The area A_z under the ROC curve versus the number of features, k, used in linear discriminant analysis for Case 1 (identity covariance matrix). In this figure, it is assumed that an infinite number of features are available for classifier training, and that features $i=1,2,...,k$ are used for classification.

Case 2: Comparison of correlated and diagonal covariance matrices

Case 2(a). In this simulation condition, the number of available features was fixed at $M=100$. In contrast to the simulation condition shown in Case 1 in this section, some of the features were assumed to have non-zero correlation. The covariance matrix Σ for the 100 features was assumed to have a block-diagonal structure

$$\Sigma=\begin{bmatrix} A & 0 & 0 & \cdots & 0 \\ 0 & A & 0 & \cdots & 0 \\ 0 & 0 & A & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & A \end{bmatrix}, \tag{12}$$

where the 10-by-10 matrix A was defined as

$$A=\begin{bmatrix} 1 & 0.8 & 0.8 & \cdots & 0.8 \\ 0.8 & 1 & 0.6 & \cdots & 0.6 \\ 0.8 & 0.6 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0.6 \\ 0.8 & 0.6 & \cdots & 0.6 & 1 \end{bmatrix}, \tag{13}$$

and $\Delta\mu(i)=0.1732$ for all i. Using Eq. (2), the squared Mahalanobis distance is computed as $\Delta(\infty)=3.0$ and $A_z(\infty)=0.89$.

Case 2(b). The features given in Case 2(a) can be transformed into a set of uncorrelated features using a linear transformation, which is called the orthogonalization transformation. The linear orthogonalization transformation is defined by the eigenvector matrix of Σ, so that the covariance matrix after orthogonalization is diagonal. After the transformation, the new covariance matrix is the identity matrix, and the new mean difference vector is

$$\Delta\mu(i)=\begin{cases} 0.5477 & \text{if } i \text{ is a multiple of 10} \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

Since a linear transformation will not affect the separability of the two classes, the squared Mahalanobis distance is the same as in Case 2(a), i.e., $\Delta(\infty)=3.0$ and $A_z(\infty)=0.89$.

In practice, given a finite set of samples with correlated features, the transformation matrix to diagonalize the feature space is not known, and has to be estimated from the given samples. In our simulation study, this transformation matrix was estimated from the samples used for feature selection.

Case 3: Simulation of a possible condition in CAD

In order to simulate covariance matrices and mean vectors that one may encounter in CAD, we used texture features extracted from patient mammograms in our earlier study, which aimed at classifying regions of interest (ROIs) containing masses on mammograms as malignant or benign. Ten different spatial gray level dependence (SGLD) features were extracted from each ROI at five different distances and two directions. The number of available features was therefore $M=100$. The image processing methods that were applied to the ROI before feature extraction, and the definition of SGLD features can be found in the literature.[11,34] The

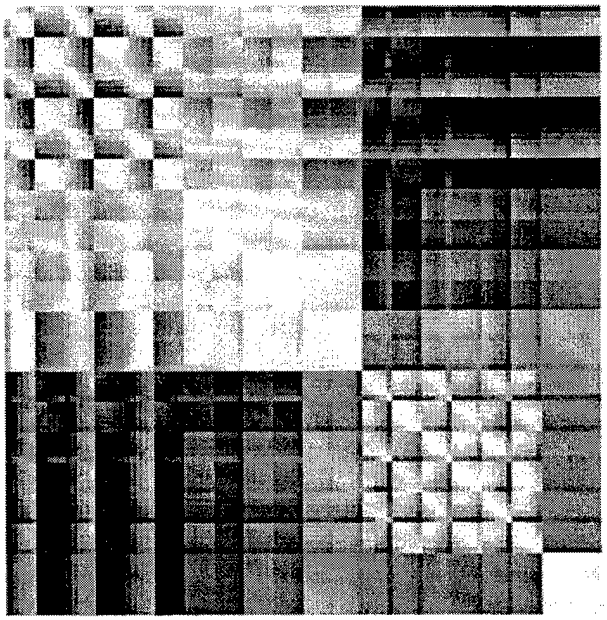


FIG. 2. The correlation matrix for the 100-dimensional texture feature space extracted from 249 mammograms. The covariance matrix corresponding to these features was used for simulations for Case 3(a).

means and covariance matrices for each class were estimated from a database of 249 mammograms. In this study, we assumed that these estimated means and covariance matrices were the true means and covariance matrices from multivariate Gaussian distribution of the population. These distributions were then used to generate random samples for the simulation study.

Case 3(a). In this simulation condition, the two classes were assumed to have a multivariate Gaussian distribution with $\Sigma=(\Sigma_1+\Sigma_2)/2$, where Σ_1 and Σ_2 were estimated from the feature samples for the malignant and benign classes. Since the feature values have different scales, their variances can vary by as much as a factor of 10^6. Therefore, it is difficult to provide an idea about how the covariance matrix looks without listing all the entries of the 100-by-100 matrix Σ. The correlation matrix, which is normalized so that all diagonal entries are unity, is better suited for this purpose. The absolute value of the correlation matrix is shown as an image in Fig. 2. In this image, small elements of the correlation matrix are displayed as darker pixels, and the diagonal elements, which are unity, are displayed as brighter pixels. From Fig. 2, it is observed that some of the features are highly correlated or anticorrelated. The squared Mahalanobis distance was computed as $\Delta(\infty)=2.4$, which corresponded to $A_z(\infty)=0.86$.

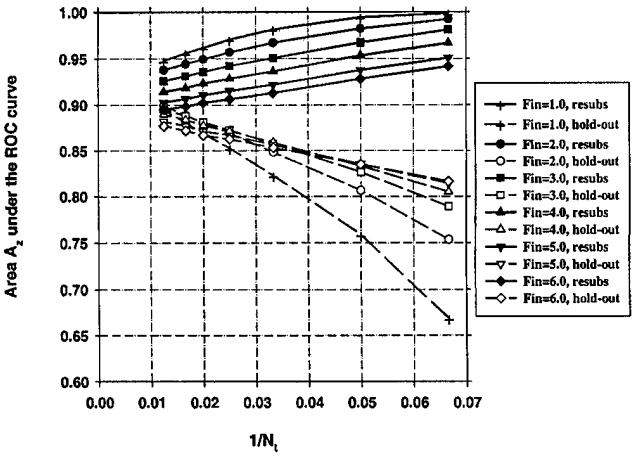


FIG. 3. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the entire sample space of 100 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M=50$ available features.

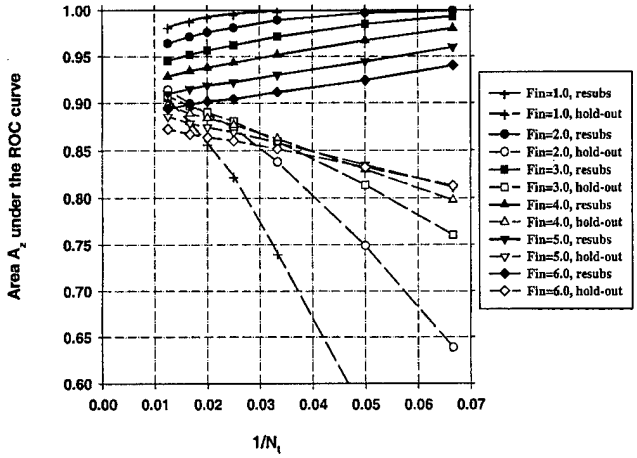


FIG. 4. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the entire sample space of 100 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M=100$ available features.

Case 3(b). To determine the performance of a feature space with equivalent discrimination potential to that in Case 3(a) but with independent features, we performed an orthogonalization transformation on the SGLD features of the generated random samples used for each partitioning, as explained previously in Case 2(b).

III. RESULTS

A. Case 1: Identity covariance matrix

1. Feature selection from entire sample space

The area A_z under the ROC curve for the resubstitution and the hold-out methods is plotted as a function of $1/N_t$ in Fig. 3 for $N_s=100$ (number of samples per class) and $M=50$ (number of available features). In this figure, the F_{in} value in stepwise feature selection is varied between 1 and 6, and $F_{\text{out}}=F_{\text{in}}-1$. Figures 4 and 5 depict the relationship between A_z and $1/N_t$ for $M=100$ and $M=200$, respectively, and $N_s=100$ for both cases. The average number of selected features for different values of F_{in} is plotted in Fig. 6. The fraction of experiments (out of a total of 50 experiments) in which feature i was selected in stepwise feature selection is plotted in Fig. 7. For the results shown in Figs. 3–7, 100 samples per class (N_s) were used in the simulation study, and the number of available features was changed from M

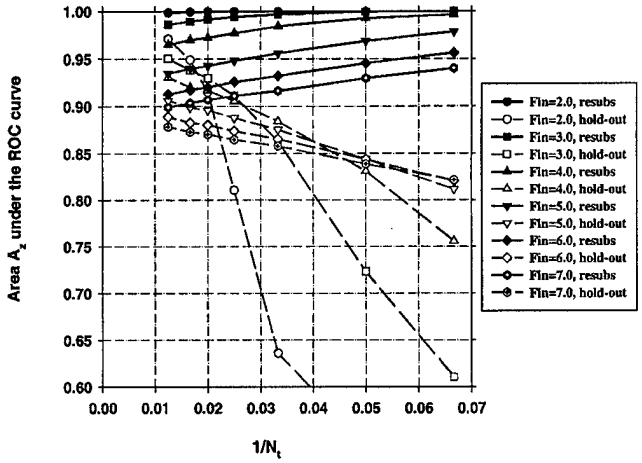


FIG. 5. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the entire sample space of 100 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M=200$ available features.

$=50$ to $M=200$. In Fig. 8, we show the simulation results for a larger number of samples, $N_s=250$, and $M=50$.

2. Feature selection from training samples alone

The area A_z under the ROC curve versus $1/N_t$ is plotted in Figs. 9–11 for $M=50$, 100, and 200, respectively. In these experiments, the number of samples per class was $N_s=100$. The average number of selected features changes as one moves along the abscissa of these curves. Figure 12 shows the average number of selected features for $N_t=80$ per class.

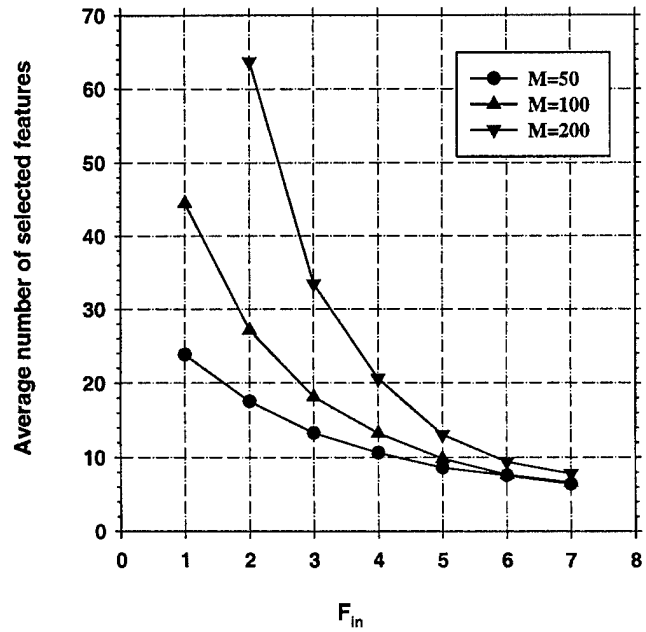


FIG. 6. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the entire sample space of 100 samples/class: The number of features selected in stepwise feature selection versus $F_{in}(F_{out}=F_{in}-1)$.

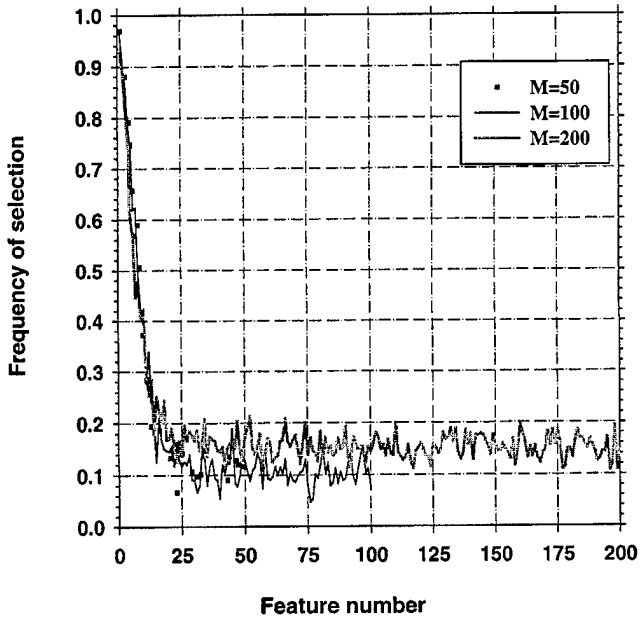


FIG. 7. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the entire sample space of 100 samples/class: The frequency of feature number i, defined as the fraction of experiments in which feature i was selected. $F_{in}=3.0$, $F_{out}=2.0$.

B. Case 2: Comparison of correlated and diagonal covariance matrices

1. Feature selection from entire sample space

The area A_z under the ROC curve for the resubstitution and hold-out methods is plotted versus $1/N_t$ in Figs. 13(a) and 13(b) for Cases 2(a) and 2(b), respectively, as described in Sec. II D for $N_s=100$ and $M=100$. Since the individual features in Case 2(a) provide less discriminatory power than those in Case 1, the F_{in} value was varied between 0.5 and 1.5 in Fig. 13(a). F_{out} was defined as $F_{out}=\max[(F_{in}-1),0]$.

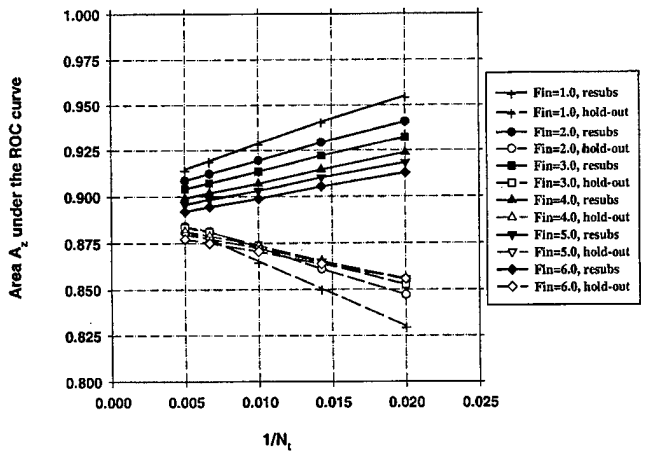


FIG. 8. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the entire sample space of 250 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M=50$ available features.

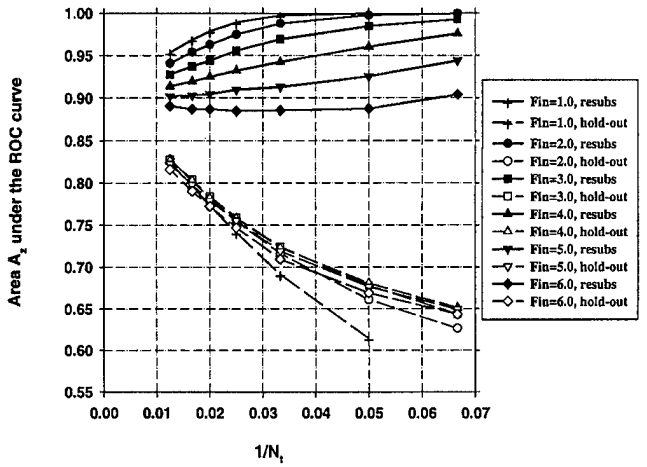


FIG. 9. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the design samples. Total sample size $N_s=100$ samples per class. The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M=50$ available features.

Figures 14(a) and 14(b) are the counterparts of Figs. 13(a) and 13(b), respectively, simulated with the number of samples per class $N_s=500$.

C. Case 3: Simulation of a possible condition in CAD

1. Feature selection from entire sample space

The area A_z under the ROC curve for the resubstitution and hold-out methods is plotted versus $1/N_t$ in Figs. 15(a) and 15(b) for Cases 3(a), and 3(b), respectively ($N_s=100$ and $M=100$). The F_{in} value was varied between 0.5 and 3.0, and F_{out} was defined as $F_{\text{out}}=\max[(F_{\text{in}}-1),0]$. Figures 16(a) and 16(b) are the counterparts of Figs. 15(a) and 15(b), simulated with the number of samples per class $N_s=500$.

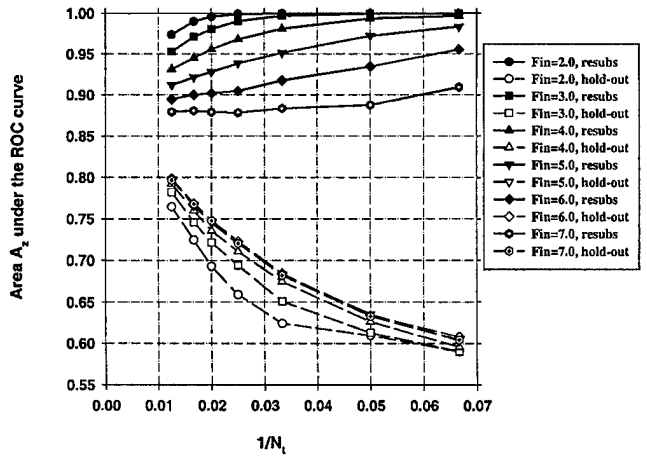


FIG. 10. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the design samples. Total sample size $N_s=100$ samples per class. The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M=100$ available features.

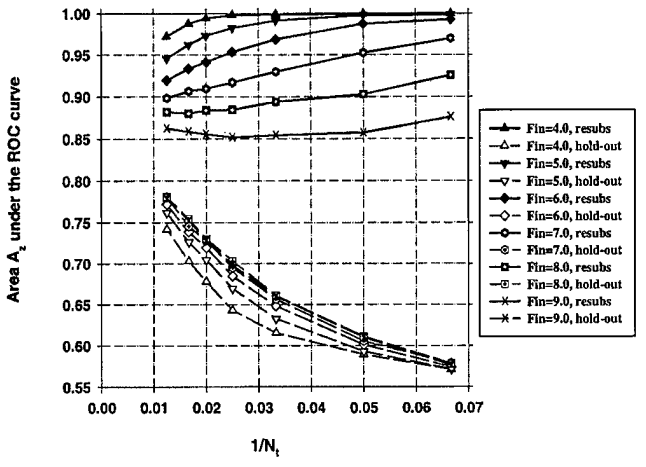


FIG. 11. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from the design samples. Total sample size $N_s=100$ samples per class. The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M=200$ available features.

2. Feature selection from training samples alone

The area A_z under the ROC curve versus $1/N_t$ for Case 3(a) is plotted for $N_s=100$ and $N_s=500$ in Figs. 17 and 18, respectively.

IV. DISCUSSION

Figures 3–5 demonstrate that, in general, when the number of available samples is fixed, the bias in the mean resubstitution performance of the classifiers increases when the number of available features increases, or when the number of selected features increases. The results also reveal the po-

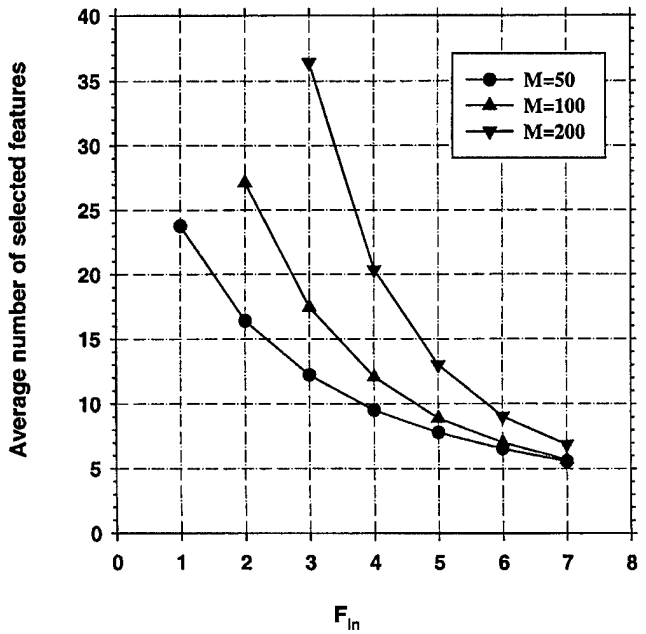


FIG. 12. Case 1 (identity covariance matrix), $A_z(\infty)=0.89$. Feature selection from $N_t=80$ design samples per class. Total sample size $N_s=100$ samples per class. The number of features selected in stepwise feature selection versus $F_{\text{in}}(F_{\text{out}}=F_{\text{in}}-1)$.

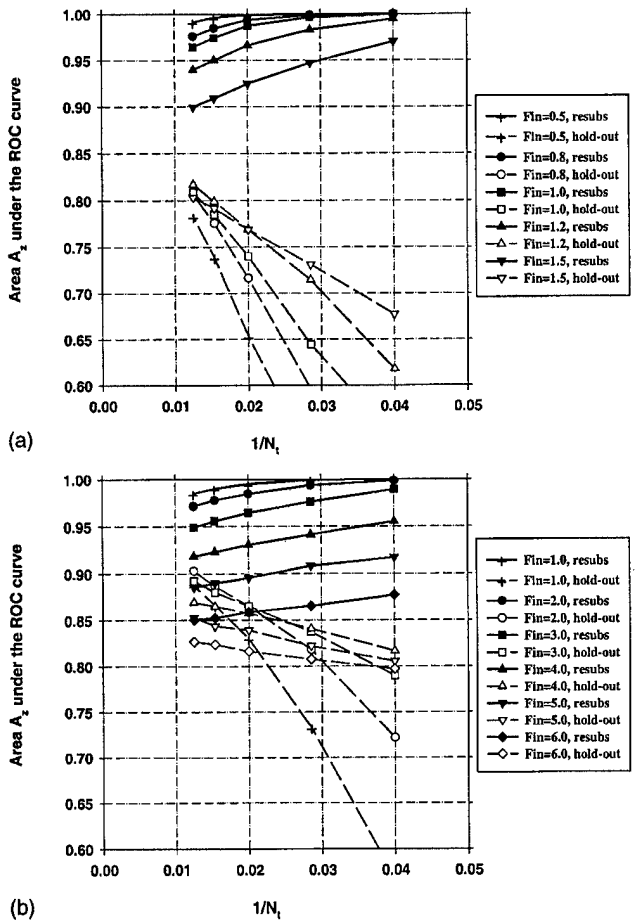


FIG. 13. (a) Case 2(a) (correlated samples, no diagonalization), $A_z(\infty) = 0.89$. Feature selection from the entire sample space of 100 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features. (b) Case 2(b) (correlated samples, and diagonalization), $A_z(\infty) = 0.89$. Feature selection from the entire sample space of 100 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features.

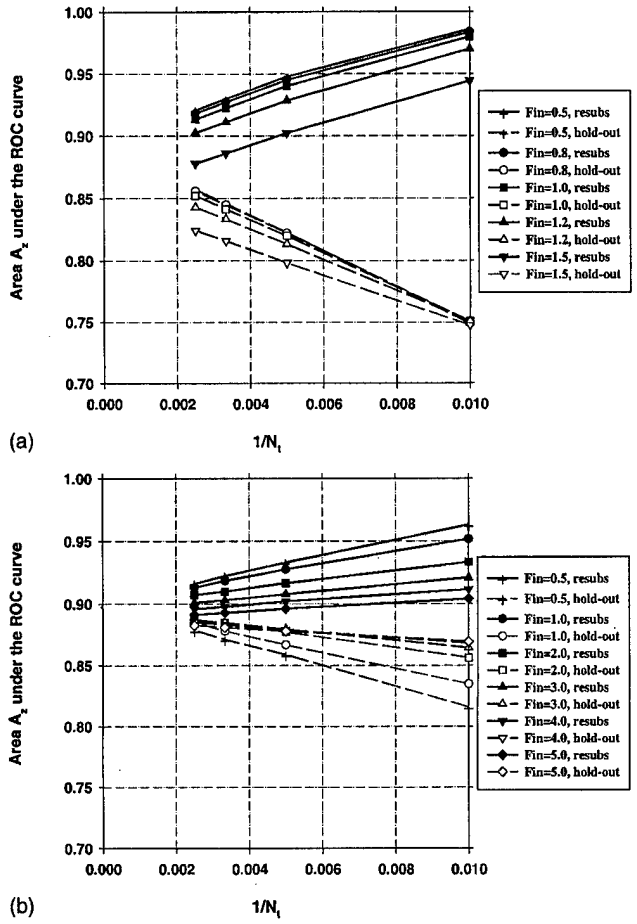


FIG. 14. (a) Case 2(a) (correlated samples, no diagonalization), $A_z(\infty) = 0.89$. Feature selection from the entire sample space of 500 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features. (b) Case 2(b) (correlated samples, and diagonalization), $A_z(\infty) = 0.89$. Feature selection from the entire sample space of 500 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features.

tential problems with the hold-out performance when feature selection is performed using the entire sample space. The best possible hold-out performance with infinite sample size for Case 1 is $A_z(\infty) = 0.89$. However, in Figs. 3–5, we observe that the "hold-out" estimates for large N_t values are higher than 0.89. Some of these estimates were as high as 0.97, as observed from Fig. 5. These hold-out A_z values were higher than $A_z(\infty)$ because the hold-out samples were not excluded from classifier design in the feature selection stage, but were excluded only in the second stage of classifier design, where the coefficients of the linear classifier were computed. When feature selection is performed using a small sample size, some features that are useless for the general population may appear to be useful for the classification of the small number of samples at hand. This was previously demonstrated in the literature[35] by comparing the probability of misclassification based on a finite sample to that based on the entire population when a certain number of features were used for classification. In our study, given a small data set, the variance in the Wilks' lambda estimates causes some feature combinations to appear more powerful than they actually are. Recall that for Case 1, the discriminatory power of a given feature decreases with the feature number. Figure 7 demonstrates that the features numbered larger than 100, which have practically no classification capability, have more than 10% chance of being selected when $F_{\text{in}} = 3.0$ and $F_{\text{out}} = 2.0$. If training-test partitioning is performed after feature selection, and a relatively large portion of the available samples are used for training so that the estimation of linear discriminant coefficients is relatively accurate, the hold-out estimates can be optimistically biased. Figures 3–5 suggest that a larger dimensionality of the available feature space (M) may imply a larger bias. This is expected intuitively because, by using a larger number of features, one increases the chance of finding a feature that is useless but appears to be useful due to a finite sample size.

The observation made in the previous paragraph about the possible optimistic bias of the hold-out estimate when fea-

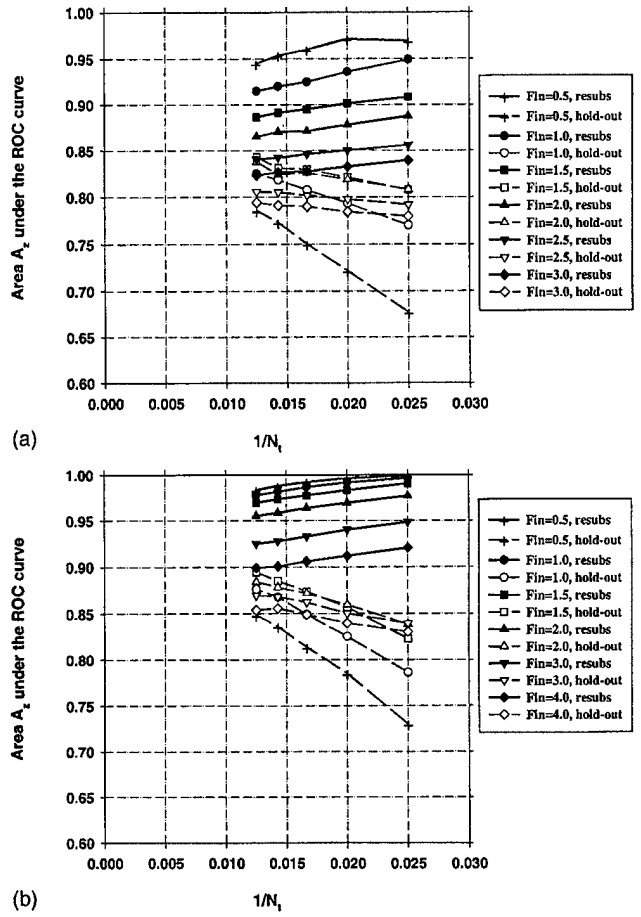


FIG. 15. (a) Case 3(a) (an example from CAD, no diagonalization), $A_z(\infty) = 0.86$. Feature selection from the entire sample space of 100 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features. (b) Case 3(b) (an example from CAD, and diagonalization), $A_z(\infty) = 0.86$. Feature selection from the entire sample space of 100 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features.

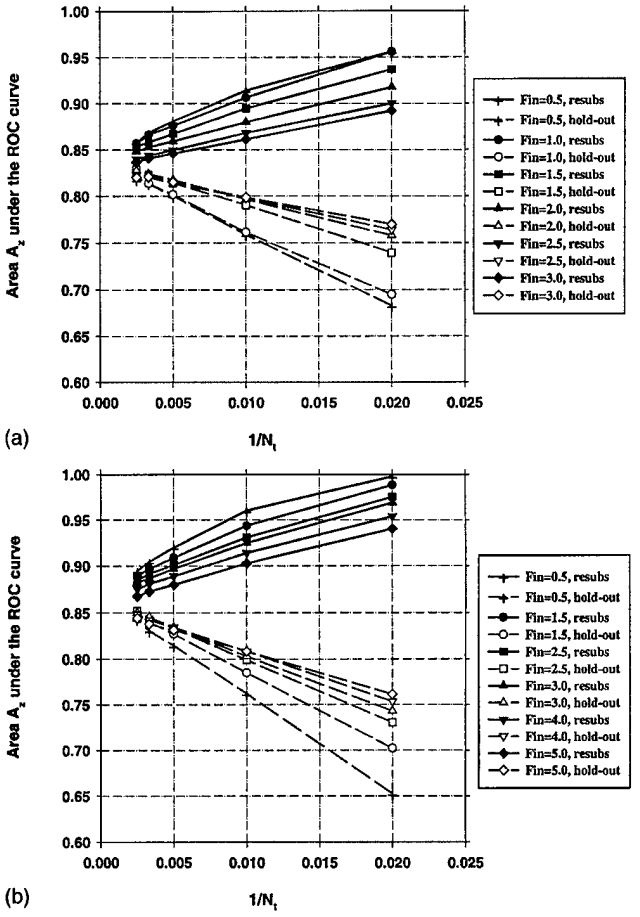


FIG. 16. (a) Case 3(a) (an example from CAD, no diagonalization), $A_z(\infty) = 0.86$. Feature selection from the entire sample space of 500 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features. (b) Case 3(b) (an example from CAD, and diagonalization), $A_z(\infty) = 0.86$. Feature selection from the entire sample space of 500 samples/class: The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features.

ture selection is performed using the entire sample space is not a general rule. Figures 13(a) and 15(a) show that one does not always run the risk of obtaining an optimistic bias in the hold-out estimate when the feature selection is performed using the entire sample space, even when the size of the entire sample space is small ($N_s = 100$) and the dimensionality of the feature space is large ($M = 100$). For Case 2, the best possible test performance with infinite sample size is $A_z(\infty) = 0.89$, however, the best hold-out estimate in Fig. 13(a) is $A_z = 0.82$. Similarly, for Case 3, the best possible test performance with infinite sample size is $A_z(\infty) = 0.86$, but the best hold-out estimate in Fig. 15(a) is $A_z = 0.84$. The features in both Cases 2(a) and 3(a) were correlated. Cases 2(b) and 3(b) were obtained from Cases 2(a) and 3(a) by applying a linear orthogonalization transformation to the features so that they become uncorrelated. Note that the linear transformation matrix is estimated from the samples used for feature selection, so it can be considered to be part of the feature selection process. Figures 13(b) and 15(b) show that after this transformation is applied, the hold-out estimates can be optimistically biased for small sample size ($N_s = 100$). However, in the range of small training sample size (N_t) below about 50, the orthogonalization reduces the biases and thus improves the performance estimation. This shows that performing a linear combination of features before stepwise feature selection can have a strong influence on its performance. This result is somewhat surprising, because the stepwise procedure is supposed to select a set of features whose linear combination can effectively separate the classes. One possible reason is that the orthogonalization transformation is applied to the entire feature space of M features, whereas the stepwise procedure only produces combinations of a subset of these features.

Figures 9–11, 17, and 18 demonstrate that, when feature selection is performed using the training set alone, the hold-out performance estimate is pessimistically biased. The bias increases, as expected, when the number of available features is increased from $M = 50$ in Fig. 9 to $M = 200$ in Fig. 11.

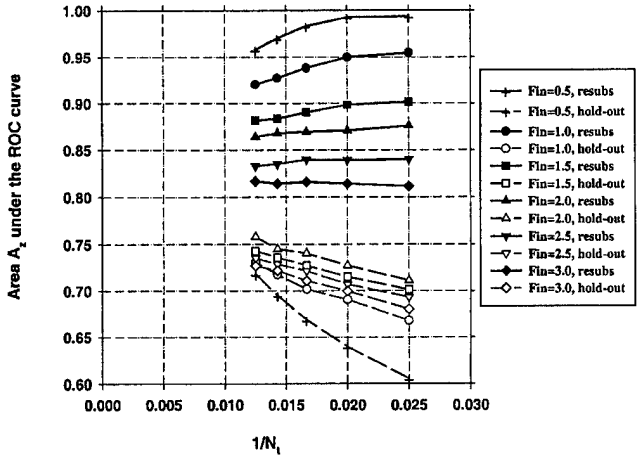


FIG. 17. Case 3(a) (an example from CAD, no diagonalization), $A_z(\infty) = 0.86$. Feature selection from the design samples. Total sample size $N_s = 100$ samples per class. The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features.

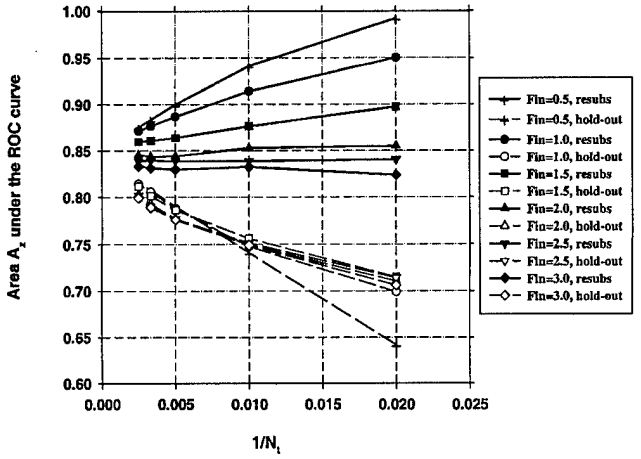


FIG. 18. Case 3(a) (an example from CAD, no diagonalization), $A_z(\infty) = 0.86$. Feature selection from the design samples. Total sample size $N_s = 500$ samples per class. The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class. Feature selection was performed using an input feature space of $M = 100$ available features.

When a larger number of features are available, it is more likely that there will be features that appear to be more useful for the classification of training samples than they actually are for the general population. This bias reduces as the number of training samples, N_t, increases.

The biases of the hold-out performance estimates discussed above are summarized in Table I when the number of available features $M = 100$. When $N_s = 100$, Cases 1, 2(b), and 3(b) can exhibit optimistic hold-out estimates if the feature selection is performed using the entire sample space. When the number of available samples is increased to $N_s = 500$, we do not observe this undesired behavior, and all the hold-out performance estimates are conservative. When the feature selection is performed using the training set alone, the average hold-out performance estimate is always pessimistically biased.

Figure 6 plots the number of selected features for Case 1 versus the F_{in} value when feature selection is performed using the entire sample space of 100 samples per class. It is observed that, for a given F_{in} value, the number of selected features increases when the number of available features M is increased. Figure 12 shows a similar trend between the number of selected features, the F_{in} value, and the number of available features when feature selection is performed using the training set alone.

When the F_{in} and F_{out} values were low, the resubstitution performance estimates were optimistically biased for all the cases studied. Low F_{in} and F_{out} values imply that many features are selected using the stepwise procedure. From previous studies, it is known that a larger number of features in classification implies larger resubstitution bias.[1,3] On the other hand, when F_{in} and F_{out} values were too high, the number of selected features could be so low that even the resubstitution estimate would be pessimistically biased, as can be observed from Fig. 14(a) ($F_{\text{in}} = 1.5$) and Fig. 15(a) ($F_{\text{in}} = 3.0$). In all of our simulations, for a given number of training samples N_t, the resubstitution estimate increased monotonically as the number of selected features were increased by decreasing F_{in} and F_{out}.

In contrast to the resubstitution estimate, the hold-out estimate for a given number of training samples did not change monotonically as F_{in} and F_{out} were decreased. This trend is apparent in Fig. 4, where the hold-out estimate at $N_t = 80 (1/N_t = 0.0125)$ is the largest for $F_{\text{in}} = 2.0$, but at N_t

TABLE I. Summary of the hold-out performance bias with respect to infinite sample performance for the class distributions used in this study. Number of available samples $M = 100$. P: Always pessimistically biased for all F_{in} and F_{out} thresholds used in stepwise feature selection in this study; O: Could be optimistically biased for some F_{in} and F_{out} thresholds used in stepwise feature selection.

	Samples per class	Case 1	Case 2(a)	Case 2(b)	Case 3(a)	Case 3(b)
Feature selection from the entire sample space	$N_s = 100$	O	P	O	P	O
	$N_s = 500$	P	P	P	P	P
Feature selection from the design samples alone	$N_s = 100$	P	P	P	P	P

$=30(1/N_t=0.033)$ it is next-to-smallest for the same F_{in} value. Another way of examining the same phenomenon is to consider different $1/N_t$ values on the abscissa of Fig. 4, and to observe that at different $1/N_t$ values, a different F_{in} threshold provided the best hold-out performance. In Fig. 4, the feature selection was performed using the entire sample space. A similar phenomenon can be observed in Fig. 18, where the feature selection is performed using the training samples alone. This means that for a given number of design samples, there is an optimal value for F_{in} and F_{out} (or number of selected features) that provides the highest hold-out estimate. This is the well-known peaking phenomenon described in the literature.[36] For a given number of training samples, increasing the number of features in the classification has two opposing effects on the hold-out performance. On the one hand, the new features may provide some new information about the two classes, which tends to increase the hold-out performance. On the other hand, the increased number of features increases the complexity of the classifier, which tends to decrease the hold-out performance. Depending on the balance between how much new information the new features provide and how much the complexity increases, the hold-out performance may increase or decrease when the number of features is increased.

For different cases studied here, the range of F_{in} and F_{out} values shown in the performance-versus-$1/N_t$ plots was different. As mentioned in the Methods Section, F_{in} and F_{out} values for a given classification task are not known *a priori*, and these thresholds have to be varied over a range in order to find the best combinations of features. As mentioned in the previous paragraph, for a given number of design samples, there is an optimum value for F_{in} and F_{out} that provides the highest hold-out estimate. In this study, we aimed at finding this peak for the highest N_t in a given graph whenever possible. After this peak was found, the F_{in} and F_{out} values shown in the figures were chosen to demonstrate the performance of the classifier at each side of the peak. By examining the figures, it can be observed that the peak hold-out performance was found in every case except in Fig. 5. In Fig. 5, the best hold-out performance occurs for $F_{in}=2.0$, for which the resubstitution performance is 1.0 for all N_t values, and the hold-out performance is 0.97. Since this F_{in} value already shows that the hold-out performance can be too optimistic, we did not search further for the peak of the hold-out performance in Fig. 5.

An interesting observation is made by examining the resubstitution performances in Figs. 9, 17, and 18, in which the feature selection is performed using the design samples alone. For $F_{in}=6.0$ in Fig. 9, and $F_{in}=3.0$ in Figs. 17 and 18, the resubstitution estimate increases as the number of training samples N_t increases. This may seem to contradict some previous studies in which the resubstitution estimate always decreased with increasing N_t.[2] However, Figs. 9, 17, and 18 are different from previous studies in that the number of selected features changes as N_t changes in these figures. The number of features selected by the stepwise procedure depends on the number of samples used for selection, which is

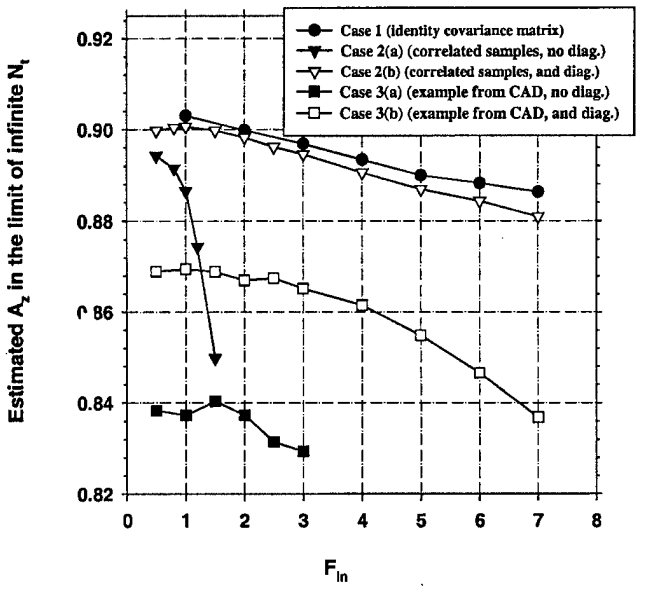


FIG. 19. The estimated values of classifier accuracy in the limit of infinite training samples, obtained by fitting a linear regression to the hold-out A_z values, and finding the y-axis intercept. $A_z(\infty)=0.89$ for Cases 1, and 2; $A_z(\infty)=0.86$ for Case 3. For all cases, total sample size $N_s=500$ samples per class, and number of available features $M=100$.

equal to $2N_t$ in these figures. With an argument similar to that for the hold-out performance, there are two opposing factors that affect the resubstitution performance when N_t is increased. The first factor, which seems to be dominant, is the fact that, with large N_t, overtraining is decreased so that the resubstitution performance is reduced. The second factor, which is visible for $F_{in}=6.0$ in Fig. 9, and $F_{in}=3.0$ in Figs. 17 and 18, is the fact that with large N_t, the stepwise procedure selects more features, which may increase the resubstitution performance.

In this study, for Cases 1, 2, and 3, we investigated the classifier performance when feature selection was performed using the entire sample space, and the number of samples per class (N_s) was five times that of available features for feature selection (M). The results of these simulations are shown in Figs. 8, 14, and 16 for Cases 1, 2 and 3, respectively. Our first observation concerning these figures is that none of the hold-out estimates in these figures are higher than their respective $A_z(\infty)$ values. This suggests that it may be possible to avoid obtaining optimistic hold-out estimates by increasing the number of available samples or by decreasing the number of features used for feature selection. A second observation is that, compared to other results in this study, the relationship between the A_z values and $1/N_t$ is closer to a linear relation in these figures. In order to test whether the $A_z(\infty)$ value can be obtained by extrapolation as was suggested in the literature,[2,17] we performed regression analysis for the hold-out A_z estimates (versus $1/N_t$) for each F_{in} value, and computed the y-axis intercept of the resulting regression equation. For regression analysis, we used curves obtained with $N_s=500$ and $M=100$ for all cases (shown in Figs. 14 and 16 for Cases 2 and 3, and not shown for Case 1). The resulting extrapolated values are shown in Fig. 19.

For Case 1, we observe that the extrapolated value is within ± 0.015 of the $A_z(\infty)$ value of 0.89. For Cases 2 and 3, the extrapolated values are within ± 0.02 of the $A_z(\infty)$ values for small F_{in}; the error increases, however, when F_{in} is increased. This graph suggests that when the classifier design involves feature selection, it may be possible to estimate the $A_z(\infty)$ value using the Fukunaga–Hayes method when the sample size is reasonably large. However, the error in the estimated $A_z(\infty)$ value can be large if the F_{in} and F_{out} thresholds are not chosen properly.

This study examined only the bias of the mean performance estimates, which were obtained by averaging the estimates from fifty experiments as described in Sec. II D. Another important issue in classifier design and assessment is the uncertainty in the performance measure, i.e., the variance expected over replications of the experiment when a new sample of training patients and/or a new sample of test patients are drawn from the same population. The variance provides an estimate of the generalizability of the classifier performance to other design and test samples. We previously studied the components of the variance of performance estimates when the classifier is trained and tested with finite samples, but the design excludes the feature selection process.[4,5] The extension of our previous studies to include feature selection is an important further research topic.

V. CONCLUSION

In this study, we investigated the finite-sample effects on the mean performance of a linear classifier that included stepwise feature selection as a design step. We compared the resubstitution and hold-out estimates to the true classification accuracy, which is the performance of a classifier designed with the full knowledge of the population distributions. We compared the effect of partitioning the data set into training and test groups before performing feature selection with that after performing feature selection. When data partitioning was performed before feature selection, the hold-out estimate was always pessimistically biased. When partitioning was performed after feature selection, i.e., the entire sample space was used for feature selection, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, number of available samples, and their statistical distribution. All hold-out estimates exhibited a pessimistic bias when the parameters of the simulation were obtained from correlated texture features extracted from mammograms in our previous study. The understanding of the performance of the classifier designed with different schemes will allow us to utilize a limited sample set efficiently and to avoid an overly optimistic assessment of the classifier.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 48129, by a Career Development Award (B.S.) from the USAMRMC (DAMD 17-96-1-6012), and a Whitaker Foundation Grant (N.P.). The authors are grateful to Charles E. Metz, Ph.D., for providing the CLABROC program.

[a]Author to whom correspondence should be addressed. Telephone: (734)647-7429; Fax: (734)647-8557. Electronic mail: berki@umich.edu
[b]Center for Devices and Radiological Health, FDA, Rockville, Maryland 20857.

[1] H.-P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: Quadratic and neural network classifiers," Proc. SPIE Conf. Medical Imaging **3034**, 1102–1113 (1997).
[2] R. F. Wagner, H.-P. Chan, J. Mossoba, B. Sahiner, and N. Petrick, "Finite-sample effects and resampling plans: Application to linear classifiers in computer-aided diagnosis," Proc. SPIE Conf. Medical Imaging **3034**, 467–477 (1997).
[3] H.-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Effects of sample size on classifier design for computer-aided diagnosis," Proc. SPIE Conf. Medical Imaging **3338**, 845–858 (1998).
[4] R. F. Wagner, H.-P. Chan, J. Mossoba, B. Sahiner, and N. Petrick, "Components of variance in ROC analysis of CAD_x classifier performance," Proc. SPIE Conf. Medical Imaging **3338**, 859–875 (1998).
[5] R. F. Wagner, H.-P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Components of variance in ROC analysis of CAD_x classifier performance: Applications of the bootstrap," Proc. SPIE Conf. Medical Imaging **3661**, 523–532 (1999).
[6] H.-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," Med. Phys. **26**, 2654–2668 (1999).
[7] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans* (Society for Industrial and Applied Mathematics, Philadelphia, 1982).
[8] C.-M. Wu, Y.-C. Chen, and K.-S. Hsieh, "Texture feature for classification of ultrasonic liver images," IEEE Trans. Med. Imaging **11**, 141–152 (1992).
[9] L. M. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, and M. A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," IEEE Trans. Med. Imaging **18**, 1178–1187 (1999).
[10] P. A. Freeborough and N. C. Fox, "MR image texture analysis applied to the diagnosis and tracking of Alzheimer's disease," IEEE Trans. Med. Imaging **17**, 475–479 (1998).
[11] B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1998).
[12] B. S. Garra, B. H. Krasner, S. C. Horri, S. Ascher, S. K. Mun, and R. K. Zeman, "Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis," Ultrason. Imaging **15**, 267–285 (1993).
[13] K. G. A. Gilhuijs and M. L. Giger, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," Med. Phys. **25**, 1647–1654 (1998).
[14] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," IEEE Trans. Med. Imaging **14**, 537–547 (1995).
[15] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer," Radiology **187**, 81–87 (1993).
[16] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Med. Phys. **19**, 1475–1481 (1992).
[17] K. Fukunaga and R. R. Hayes, "Effects of sample size on classifier design," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 873–885 (1989).
[18] P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).
[19] M. M. Tatsuoka, *Multivariate Analysis, Techniques for Educational and Psychological Research*, 2nd ed. (Macmillan, New York, 1988).
[20] N. R. Draper, *Applied Regression Analysis* (Wiley, New York, 1998).
[21] H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).
[22] H.-P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic

microcalcifications in morphological and texture feature spaces," Med. Phys. **25**, 2007–2019 (1998).
[23] N. Petrick, H.-P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and tissue classification," Med. Phys. **23**, 1685–1696 (1996).
[24] M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics* (SPSS Inc., Chicago, IL, 1993).
[25] W. W. Peterson, T. G. Birdsall, and W. C. Fox, "The theory of signal detectability," Trans. IRE Prof. Grp. Inform. Theory **PGIT-4**, 171–212 (1954).
[26] W. P. Tanner and J. A. Swets, "A decision-making theory of visual detection," Psychol. Rev. **61**, 401–409 (1954).
[27] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
[28] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," Invest. Radiol. **14**, 109–121 (1979).
[29] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology **143**, 29–36 (1982).
[30] C. E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720–733 (1986).
[31] A. J. Simpson and M. J. Fitter, "What is the best index of detectability," Psychol. Bull., 80 (1973).
[32] C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Statistics in Medicine **17**, 1033–1053 (1998).
[33] D. Dorfman and E. Alf, "Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data," J. Math. Psychol. **6**, 487 (1969).
[34] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).
[35] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," IEEE Trans. Pattern Anal. Mach. Intell. **13**, 252–264 (1991).
[36] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Inf. Theory **14**, 55–63 (1968).

Original Investigations

Digital Mammography:

Observer Performance Study of the Effects of Pixel Size on the Characterization of Malignant and Benign Microcalcifications[1]

Heang-Ping Chan, PhD, Mark A. Helvie, MD, Nicholas Petrick, PhD, Berkman Sahiner, PhD, Dorit D. Adler, MD
Chintana Paramagul, MD, Marilyn A. Roubidoux, MD, Caroline E. Blane, MD, Lynn K. Joynt, MD
Todd E. Wilson, MD, Lubomir M. Hadjiiski, PhD, Mitchell M. Goodsitt, PhD

Rationale and Objectives. The authors performed this study to evaluate the effects of pixel size on the characterization of mammographic microcalcifications by radiologists.

Materials and Methods. Two-view mammograms of 112 microcalcification clusters were digitized with a laser scanner at a pixel size of 35 μm. Images with pixel sizes of 70, 105, and 140 μm were derived from the 35-μm–pixel size images by averaging neighboring pixels. The malignancy or benignity of the microcalcifications had been determined with findings at biopsy or 2-year follow-up. Region-of-interest images containing the microcalcifications were printed with a laser imager. Seven radiologists participated in a receiver operating characteristic (ROC) study to estimate the likelihood of malignancy. The classification accuracy was quantified with the area under the ROC curve (A_z). The statistical significance of the differences in the A_z values for different pixel sizes was estimated with the Dorfman-Berbaum-Metz method and the Student paired t test. The variance components were analyzed with a bootstrap method.

Results. The higher-resolution images did not result in better classification; the average A_z with a pixel size of 35 μm was lower than that with pixel sizes of 70 and 105 μm. The differences in A_z between different pixel sizes did not achieve statistical significance.

Conclusion. Pixel sizes in the range studied do not have a strong effect on radiologists' accuracy in the characterization of microcalcifications. The low specificity of the image features of microcalcifications and the large interobserver and intraobserver variabilities may have prevented small advantages in image resolution from being observed.

Key Words. Breast neoplasms, calcification; breast radiography, comparative studies; breast radiography, technology; receiver operating characteristic curve (ROC).

Breast cancer is one of the leading causes of death in women between the ages of 40 and 55 years. In the United States, the mortality rate for breast cancer in women is the second highest of all cancers, and breast cancer was estimated to account for 16% of all cancer deaths in 1998 (1). Studies have indicated that early detection and treatment improve the chances of survival for breast cancer patients. At present, mammography is the only proven method that consistently demonstrates minimal breast cancers (2,3). The image quality with conventional mammography, however, is limited by the dynamic range of screen-film systems. The contrast sensitivity of screen-film mammograms is very poor in the overpen-

Acad Radiol 2001; 8:454–466

[1] From the Department of Radiology, University of Michigan Hospital, UH B1F510, Ann Arbor, MI 48109-0030. Received September 6, 2000; revision requested October 3; revision received January 10, 2001; accepted January 11. Supported by U.S. Public Health Service grant CA 48129 and by a grant from the U.S. Army Medical Research and Materiel Command DAMD 17-96-1-6254. B.S. and L.M.H. supported by Career Development Awards DAMD 17-96-1-6012 and 17-98-1-8211, respectively. **Address correspondence to** H.P.C.

etrated periphery and the underpenetrated dense fibroglandular tissue regions on the breast image. Recently, a digital mammography system has received U.S. Food and Drug Administration clearance for clinical use. Digital mammography detectors are expected to provide a wider dynamic range than screen-film systems and, thus, increase the contrast sensitivity in the periphery and dense regions of the breast. The improved image quality is expected to lead to an improvement in the accuracy of breast cancer diagnosis.

The spatial resolution of current digital detectors is generally lower than that of screen-film systems. Digital detectors used in the full-field digital mammography systems that are commercially available or under development have pixel sizes in the range of 40 × 40 μm to 100 × 100 μm, which correspond to nominal spatial resolution of about 12 line pairs per millimeter to 5 line pairs per millimeter. In contrast, the spatial resolution of mammographic screen-film systems generally exceeds 20 line pairs per millimeter. Higher-resolution digital detectors require smaller pixel sizes. The development of digital detectors with small pixel sizes, however, is not only technologically demanding, but the requirements for image transmission, archiving, and display increase rapidly as the matrix size increases. The trade-offs between spatial resolution and cost and efficiency are important considerations in the development of digital mammography systems. The maximum pixel size acceptable for performing mammography without reducing the detectability of subtle breast cancers is unknown.

One of the important signs of breast cancer is clustered microcalcifications (4), which can be seen on mammograms in 30%–50% of breast cancers (5–8). Microcalcifications associated with early breast cancers are usually smaller than about 500 μm. Among the image features that may indicate the presence of breast cancer, microcalcifications are the smallest. Therefore, the spatial resolution required for the detection and characterization of subtle microcalcifications on mammograms may be regarded as the lower bound for the resolution of a mammographic detector. In a previous receiver operating characteristic (ROC) study (9), we compared the detectability of subtle microcalcifications on original screen-film mammograms with that on mammograms digitized at a pixel size of 100 μm with an optical drum scanner. We found that the detection accuracy of subtle microcalcifications decreased when radiologists read the digitized images. Although the detection accuracy improved after the digitized images were enhanced with unsharp mask filtering, it remained lower than that with the original screen-film mammograms. In another study (10), we investigated the detectability of individual microcalcifications on digitized mammograms by using a computer program. Those results also indicated a reduction in detectability when the digitization pixel size increased from 35 to 140 μm.

Malignant microcalcifications may exhibit linear and branching shapes, as well as variations in shape and size within a cluster. Benign microcalcifications tend to be round and smooth, with relatively uniform shapes and sizes within a cluster. The visibility of the detailed shapes is dependent on the spatial resolution of the image recording system. Therefore, it is generally believed that a higher spatial resolution is required to differentiate malignant from benign microcalcifications than to detect microcalcifications. Results of some recent studies, however, indicate that this may not be the case. Karssemeijer et al (11) performed an ROC study to compare the accuracy of classifying microcalcifications on original screen-film mammograms with that on images digitized at a pixel size of 100 μm and viewed on a display monitor. They found that there was no statistically significant difference in the classification accuracy between the two reading conditions. Kallergi et al (12) also performed an ROC study to compare the detection and classification of clustered microcalcifications at three reading conditions: screen-film mammograms, images digitized at a pixel size of 105 μm and displayed on a monitor, and wavelet-enhanced digitized images displayed on a monitor. They found that the detection with the original mammograms was much better than that with the digitized mammograms displayed on a monitor; the use of wavelet enhancement, however, reduced the difference. The characterization of microcalcifications was not substantially different among the three reading conditions.

We performed this ROC study to evaluate the effects of pixel size on the characterization of malignant and benign microcalcifications on digitized mammograms. Two-view mammograms were digitized and displayed as laser-printed film images at four pixel sizes ranging from 35 to 140 μm. Seven radiologists experienced in mammography estimated the likelihood of malignancy. The dependence of classification accuracy on pixel size was analyzed with ROC methodology.

MATERIALS AND METHODS

Data Set

Digital mammograms were obtained by digitizing screen-film mammograms with a laser film scanner. One

hundred twelve microcalcification clusters were selected from 100 patient cases in the Breast Imaging Division at the University of Michigan with approval from the Institutional Review Board. Two-view mammograms of each cluster were digitized. The two views included a craniocaudal view and a mediolateral oblique or lateral view.

Forty of the microcalcification clusters were proved at biopsy to be malignant, and 65 were proved at biopsy to be benign. The other seven clusters were considered to be benign based on findings of at least 2 years of follow-up. Of the 40 malignant clusters, 25 were ductal carcinoma in situ. The distribution of the sizes (the longest dimension) of the microcalcification clusters is shown in Figure 1. The longest dimension of the clusters ranged from 2.0 to 18.0 mm (mean, 6.4 mm). Seven of the benign microcalcifications and five of the malignant microcalcifications were spread over an area larger than 20 mm in diameter and, thus, were considered to be diffuse. The data set included microcalcifications with a range of subtleties. The subtlety of the microcalcifications was rated by a radiologist experienced in mammography (M.A.H.) on a scale of 1 (obvious) to 10 (subtle) relative to the visibility range of microcalcifications encountered in clinical practice. The subtlety ratings are shown in Figure 2. The malignant and benign microcalcifications were similarly distributed, with the benign microcalcifications slightly more subtle than the malignant clusters.

All mammograms were digitized at a pixel size of 35 × 35 μm with 12-bit gray levels by using a laser scanner (DIS-1000, Lumisys, Los Altos, Calif). The digitizer had an optical density range of about 0 to 3.5. It was calibrated such that the optical density on film was linearly proportional to the pixel value at 0.001 optical density units per pixel value in the optical density range of about 0–2.8. The pixel values of the images were linearly inverted so that large pixel values represented a low optical density. The resolution of the scanner was evaluated by digitizing test film images with line pair patterns. It was found that line pair patterns up to 14.3 line pairs per millimeter could be resolved on the digitized image (10).

A 1,024 × 1,024-pixel region of interest (ROI) containing the microcalcifications was extracted from the digitized image. Except for clusters that were close to the chest wall or in the breast periphery, the extracted cluster was usually centered within the ROI. Diffuse microcalcifications that were larger than the ROI were truncated to the size of the ROI. Microcalcification images digitized with pixel sizes of 70, 105, and 140 μm were simulated from the image with the 35-μm pixel size by averaging

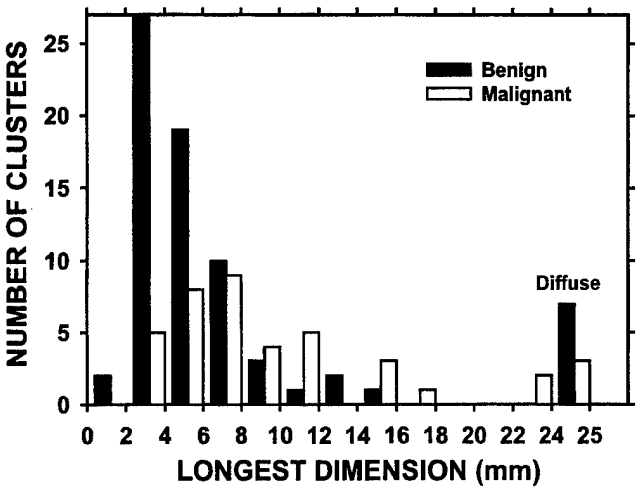


Figure 1. The size distribution of the microcalcification clusters.

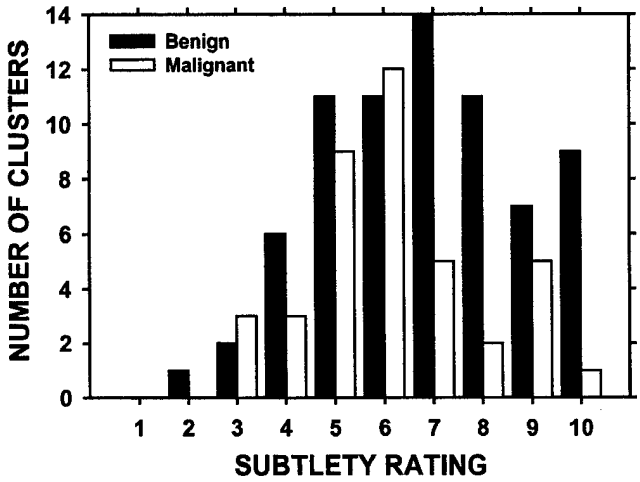


Figure 2. Distribution of the subtlety ratings for the microcalcification clusters. *1* = most obvious, *10* = most subtle.

2 × 2, 3 × 3, and 4 × 4 neighboring pixels, respectively. Because ROIs of different pixel sizes were derived from the same digitized image, there would not be differences in image quality caused by the reproducibility of digitization. The actual size of all ROIs corresponded to an area of 35.8 × 35.8 mm on the original mammograms, regardless of the pixel sizes.

Because the use of display monitors to view images can introduce variables that may be difficult to control, we printed the ROI images on film with a laser imager (model 969HQ; Imation, Oakdale, Minn) for the observer performance study. To reduce the effects of image size on characterization, the ROIs with the three larger pixel sizes (ie, smaller matrix sizes for the same ROI image) were enlarged to the same printed size as that of the 35-μm

Table 1
Confidence Rating Scale

Rating	Likelihood of Malignancy (%)	Suspicion Level	BI-RADS Category
1	0–2	Benign, probably benign	2, 3
2	3–20	Suspicious, with low probability of malignancy	4
3	21–30	Suspicious, with low probability of malignancy	4
4	31–40	Suspicious, with moderate probability of malignancy	4
5	41–50	Suspicious, with moderate probability of malignancy	4
6	51–60	Suspicious, with moderate probability of malignancy	4
7	61–70	Suspicious, with moderate probability of malignancy	4
8	71–80	Highly suggestive (high probability) of malignancy	5
9	81–90	Highly suggestive (high probability) of malignancy	5
10	91–100	Highly suggestive (high probability) of malignancy	5

pixel size images by means of interpolation. Sixteen interpolation schemes were available from the laser imager interface software. To choose the best interpolation scheme for this study, we printed an image of a cluster containing microcalcifications of different sizes and shapes at pixel sizes of 70, 105, and 140 μm by using the 16 interpolation schemes. The images of 35-μm pixel size were also printed. A radiologist who was qualified under the requirements of the Mammography Quality Standards Act visually compared the printed images and numbered his top three choices for each set of images. The radiologist was not aware of the specific schemes. After the decision was made, he informed us that his criteria were a balance between blockiness and blurriness on the enlarged image and its similarity to the 35-μm image. The experiment was repeated two times, with the sessions separated by more than a month. The top two choices obtained from the two readings were consistent. The top two choices were essentially indistinguishable so that one of them was used to print the images. The chosen scheme was a convolution interpolation that filled the interpolated pixels with smooth weighted gray levels of the adjacent pixels.

The printed ROIs measured 84 × 84 mm, which corresponded to a pixel pitch of about 82 μm for the laser imager. The printed ROIs were therefore magnified by a factor of about 2.3 compared with their size on the original screen-film mammograms. Because radiologists routinely view microcalcifications with a magnifying lens or on a magnified spot mammogram, however, the magnification should not affect the classification of the microcalcifications. To maintain the same displayed contrast for images of different pixel sizes, the four ROIs of different pixel sizes were printed on the same piece of film and, thus, developed at the same time. This minimized the effects of any potential fluctuations in the printer calibration and in the development conditions of the laser film on the relative density and contrast of the printed images.

Observer Performance Study

Seven radiologists, all of whom were qualified under the requirements of the Mammography Quality Standards Act to read and routinely interpret mammograms, participated as observers. The radiologists had 3–20 years experience in mammographic interpretation. Because there were 112 ROIs and four pixel sizes for each ROI, a total of 448 images were read by each observer. The two views of each cluster at the same pixel size were read side by side. The observers were not informed of the prevalence of malignant cases or the proportion of biopsy cases. Each observer read the ROI images in four reading sessions. Every reading session was separated from the previous one by at least 2 weeks. In each session, one-quarter of the images of each pixel size were read. Each case appeared once and only once in each session. The reading orders of the images in each pixel size were counterbalanced such that, on average, no images of a given pixel size were read in a given order (eg, read first by the observers) more often than images of any other pixel sizes. The reading order of the images was randomized differently for each observer. This systematic randomization reading scheme minimized any potential learning effects on the reading results (13). The observers were allowed unlimited reading time.

The likelihood that the microcalcifications were malignant was rated with a 10-point confidence rating scale. The confidence rating scale was designed and related to the Breast Imaging Reporting and Data System (BI-RADS) ratings by an experienced radiologist, as shown in Table 1. A likelihood of malignancy of less than 2% for benign or probably benign mammographic abnormalities was chosen on the basis of the studies by Sickles (14,15). The observers also rated the subtlety of each case according to a 10-point scale (1 = most obvious, 10 = most subtle)

on the basis of their perception of the cluster relative to their experience with clinical cases.

A table showing the rating scale and the corresponding BI-RADS category was available to the observers for reference during the reading sessions. A training session was conducted before each reading session to familiarize the observers with the rating scales. Three malignant and three benign clusters not included in the test set were used in the training session. After the rating scales were explained to the observer, he or she rated each cluster as described earlier. They were told the biopsy outcome of the cluster after rating each training case. There was no "truth" for the subtlety rating. The subtlety rating was recorded as additional information about each radiologist's subjective impression of a cluster.

Analysis of Classification Accuracy

The confidence ratings of the likelihood of malignancy were analyzed with ROC analysis (13). The two class distributions were assumed to be binormal, and an ROC curve was fitted to the confidence ratings on the basis of maximum likelihood estimation. The ROC curve represents the relationship between the true-positive fraction (sensitivity) and the false-positive fraction (1 − specificity) as the confidence threshold varies. An ROC curve was generated for each observer and for images of each pixel size. The classification accuracy was quantified by using the area under the ROC curve (A_z). The average ROC curve for each reading condition was derived by averaging the slope and intercept parameters of the individual observers' fitted ROC curves. The statistical significance of the differences in the ROC curves for two pixel sizes was estimated by using the Dorfman-Berbaum-Metz (DBM) method for multireader, multicase ROC data (16) and the Student paired t test for the observer-specific paired A_z values. The paired t test takes into account the statistical variation of the readers, whereas the DBM method includes both the reader variation and case sample variation with an analysis-of-variance approach. Therefore, the results with the DBM method can be generalized to the population of readers as well as the case samples. In addition, the bootstrap method developed by Beiden et al (17) was used to analyze the components of variances in this classification task.

RESULTS

Images of a small malignant microcalcification cluster and a benign cluster from our data set obtained with a pixel size of 35 μm are shown in Figure 3a and 3b, respectively. The craniocaudal and mediolateral oblique views of the same cluster are shown side by side. Figure 4 shows one view of a malignant cluster with all four pixel sizes. Slight blurring of the image details and the noise can be observed as the pixel size increases from 35 to 140 μm.

The ROC curves for the seven radiologists reading the images with 35-μm pixel size are shown in Figure 5. The ROC curves are spread over a relatively wide range. The A_z values for the radiologists are listed in Table 2 and plotted in Figure 6. The standard deviation of the A_z ranges from 0.05 to 0.07, as estimated with the LABMRMC program. Only one of the seven radiologists demonstrated a higher classification accuracy with the 35-μm images than with the 70- or 105-μm images. The A_z-versus–pixel size curve for this radiologist (reader 6) had a different trend from that of other radiologists. The A_z of another radiologist (reader 7) was basically constant over the entire range of pixel sizes studied. The average ROC curves for each pixel size were derived from the average slope and intercept parameters of the seven individual ROC curves and are plotted in Figure 7. The dependence of average A_z on pixel size is shown in Table 2. The average A_z showed a higher classification accuracy with pixel sizes of 70 and 105 μm. The differences in A_z between the different pixel sizes did not achieve statistical significance with either the DBM method (16) or the Student paired t test. Table 3 shows the P values obtained with the DBM and the paired t test when images with a pixel size of 35 μm were compared with those with pixel sizes of 70, 105, and 140 μm. The P values obtained with the two methods are very similar, which indicates that the reader variation is dominant over case variation in this classification task.

Because of the outlying trend of reader 6, we performed the analysis of the classification accuracy without this reader in an attempt to evaluate the dependence of A_z on pixel size for the majority of radiologists in our study. For these six readers, the average A_z for the four pixel sizes was 0.71, 0.74, 0.75, and 0.71, respectively. Although the trend that the radiologists had a higher classification accuracy with pixel sizes of 70 and 105 μm became more apparent, the difference in the A_z between the pixel sizes still fell short of statistical significance. The P value determined with the DBM method was .11 for the difference in A_z between 35- and 70-μm images and .12 for that between 35- and 105-μm images. The corre-

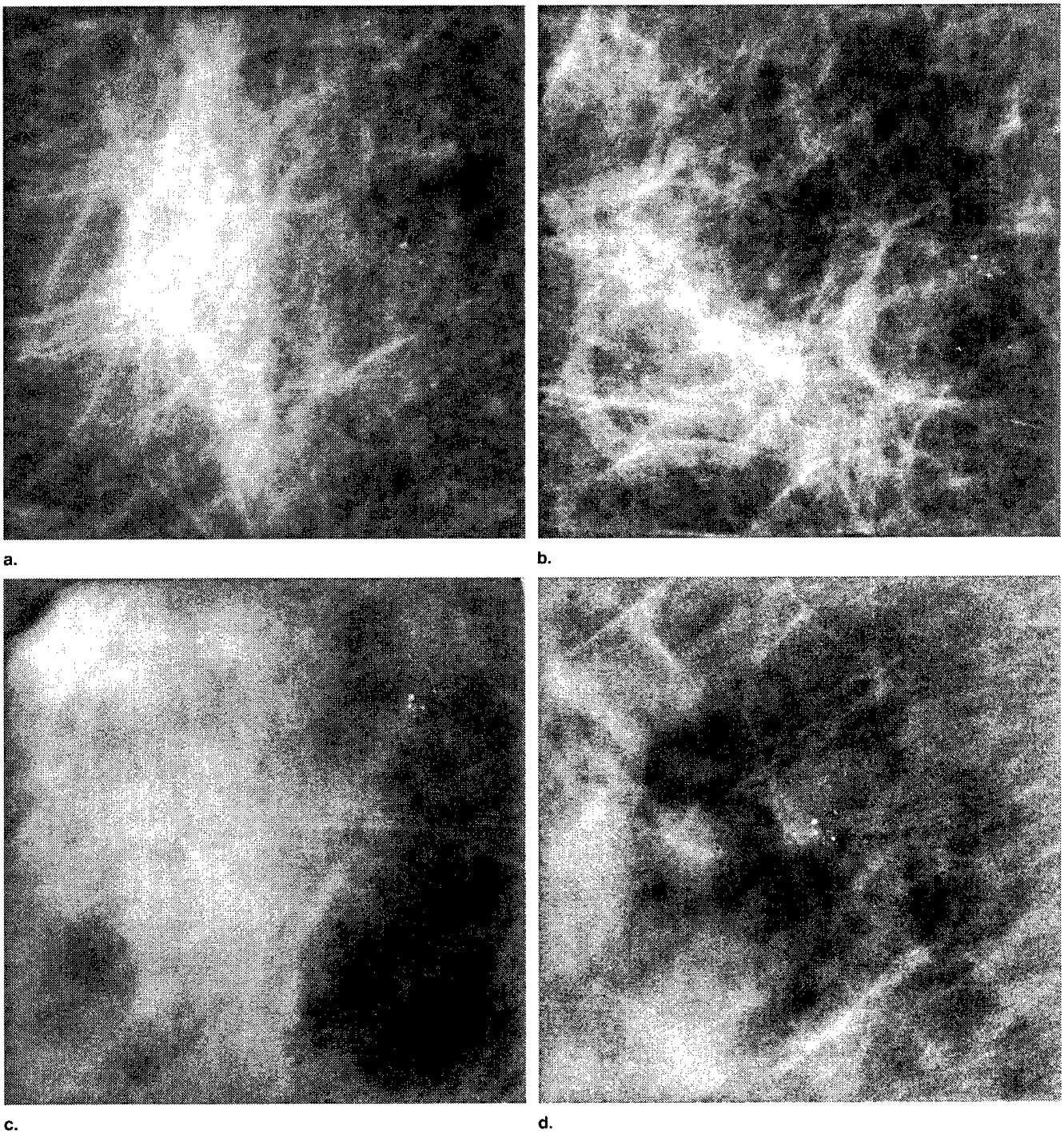


Figure 3. **(a, c)** Craniocaudal and **(b, d)** mediolateral oblique images of **(a, b)** a malignant microcalcification cluster (intraductal carcinoma) and **(c, d)** a benign cluster (sclerosing adenosis) digitized with a pixel size of 35 μm.

sponding two-tailed P values with the Student paired t test were .10 and .12, respectively.

We also analyzed the percentages of positive and negative cases for which the observers gave a confidence rating of 1 in each pixel size. A confidence rating of 1 corresponded to a 0%–2% likelihood of malignancy and BI-RADS categories of benign or probably benign (Table 1). These cases would be returned to a regular screening schedule or undergo short-interval follow-up without biopsy. The results are shown in Table 4. Each observer

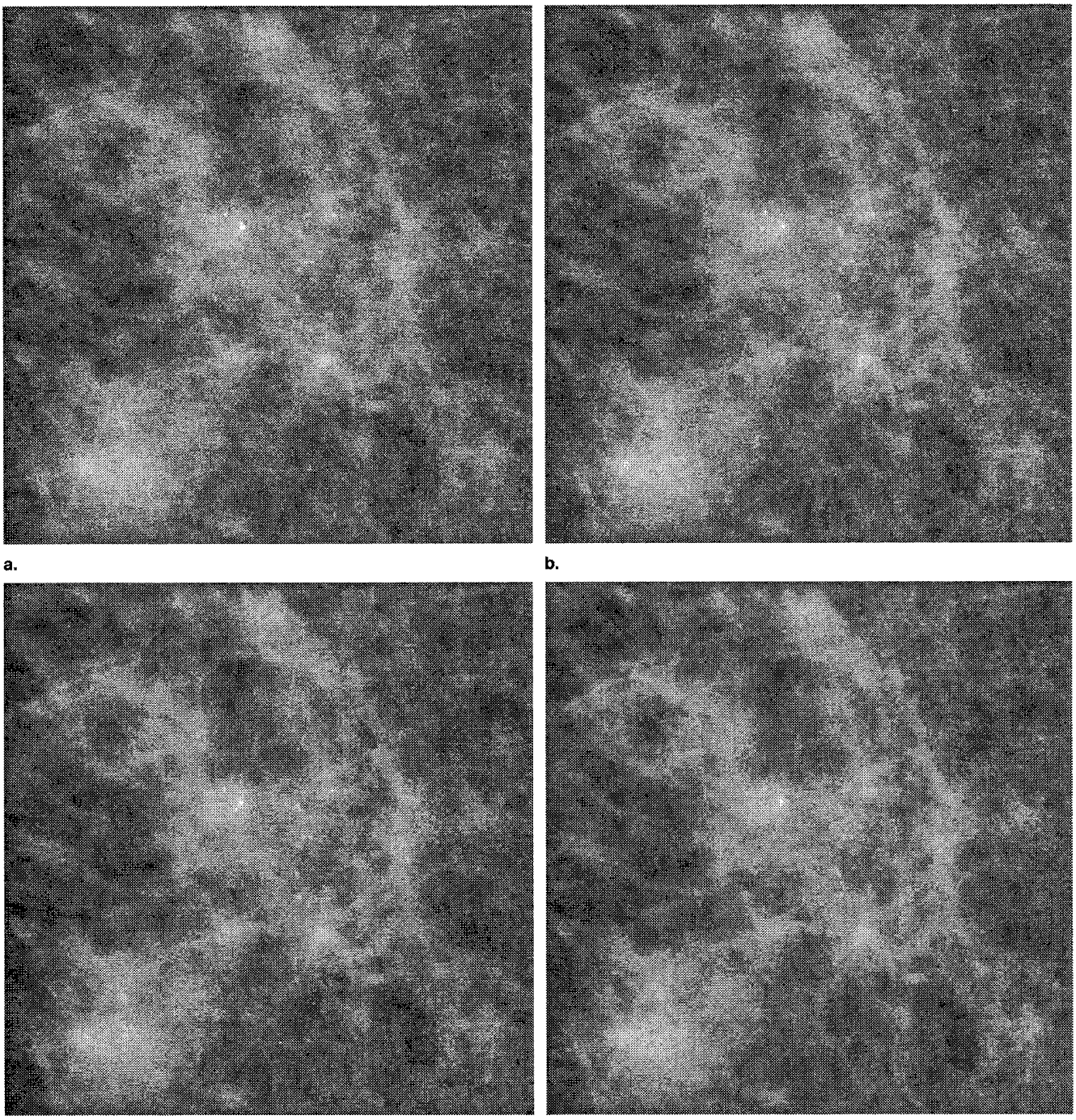


Figure 4. Lateral images of a malignant cluster (intraductal carcinoma, comedo type) at pixel sizes of **(a)** 35, **(b)** 70, **(c)** 105, and **(d)** 140 μm.

appeared to have a different threshold for suspicion. For a given observer, however, the threshold was relatively consistent among the different pixel sizes. There was no obvious trend that this threshold depended on pixel size.

DISCUSSION

The results of the ROC study indicate that the differences in the classification accuracy of microcalcifications,

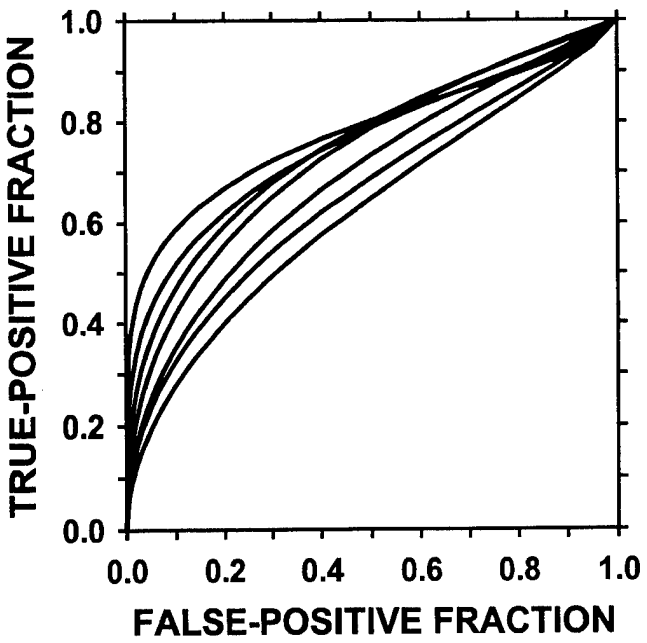


Figure 5. The ROC curves for seven radiologists in the evaluation of the images with 35-μm pixel size. The standard deviation of the A_z ranges from 0.05 to 0.07.

if any, with pixel sizes of 35–140 μm did not achieve statistical significance. Of the A_z-versus–pixel size curves from seven radiologists, only one showed that a pixel size of 35 μm provided a larger A_z than did pixel sizes of 70 and 105 μm. Although the variances in the A_z were large, this consistent trend indicates a strong likelihood that images with a 35-μm pixel size may not provide a higher accuracy in the differentiation of malignant from benign microcalcifications than those with a pixel size of 70 or 105 μm. This finding differs from the expectation that a smaller pixel size would better preserve the shape information of microcalcifications and, consequently, provide higher accuracy in the differentiation of microcalcifications on mammograms. Our findings are consistent with those of Karssemeijer et al (11) and Kallergi et al (12) who, in their ROC studies, compared the classification accuracy of microcalcifications on original screen-film mammograms with that on images digitized at a pixel size of 100 μm and viewed on a display monitor.

Beiden et al (17) recently developed a bootstrap method for analyzing the variance components in an ROC experiment. They analyzed our ROC data set and estimated the variance components and the total variance of the difference in A_z, $\sigma^2(\Delta A_z)$, for any pairing of modalities (pixel sizes), as shown in Table 5. We used these variances to estimate whether the finite sample size in our ROC study is the main factor that caused the insignificant differences between pixel sizes.

Equation (21) in the article by Beiden et al (17) shows that the total variance of ΔA_z is given as $\sigma^2(\Delta A_z) = 2[\sigma^2_{mc}(N_0/N) + \sigma^2_{mr}/R + \sigma^2_{\epsilon}(N_0/N)/R]$, where R is the number of readers; N_0 is the sample size of the current experiment; N is the sample size of a future experiment; and σ^2_{mc}, σ^2_{mr}, σ^2_{ϵ} are the modality-by-case, modality-by-reader, and effective error components of the variance, respectively. The total variance at an infinite sample size, $N\rightarrow\infty$, is thus caused only by the reader variance, as follows: $\sigma^2(N\rightarrow\infty) = 2\sigma^2_{mr}/R$. Therefore, if we can repeat the ROC experiment with an infinite sample size, the minimum observed difference in A_z between two modalities, [min $\Delta A_z(N\rightarrow\infty)$], that will allow rejection of the null hypothesis, A_z(small pixel) = A_z(large pixel), with $P < .05$ can be estimated as [min $\Delta A_z(N\rightarrow\infty)$] = $1.645 \cdot \sigma(N\rightarrow\infty)$. The values of $\sigma(N\rightarrow\infty)$ and [min $\Delta A_z(N\rightarrow\infty)$] are shown in Table 5. The z value of 1.645, which corresponds to the one-tailed P value of .05 for a normal distribution, was used in these estimations because it is expected that a smaller pixel size would provide better performance than a larger pixel size.

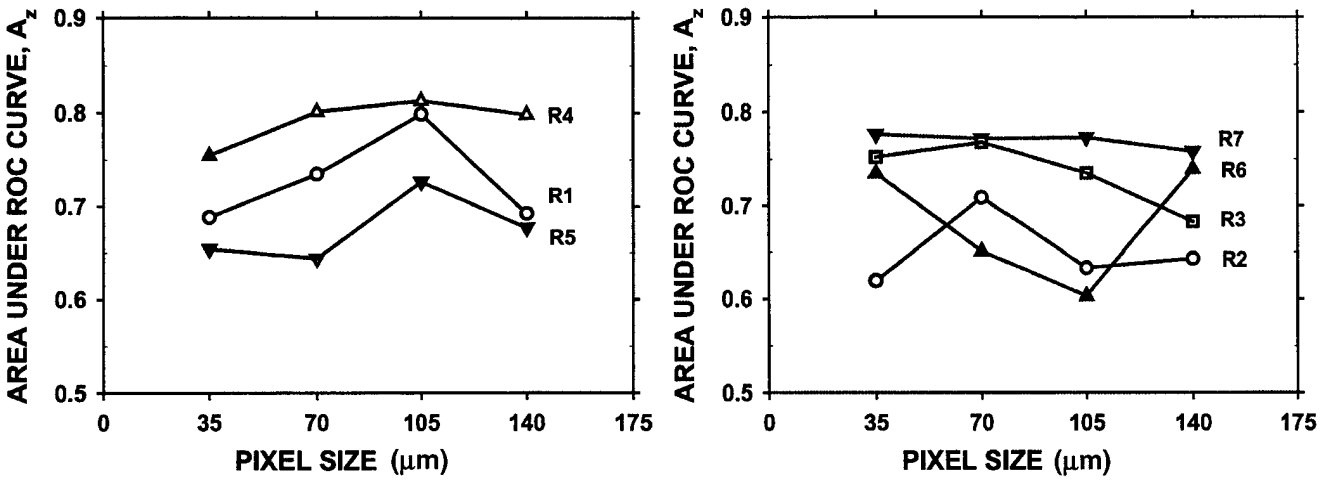
From the standard deviation, $\sigma(\Delta A_z)$, and the observed difference in A_z, we can estimate the maximum mean ΔA_z between two modalities. In our ROC experiment, we observed a difference of ΔA_z(observed) = A_z(small pixel) − A_z(large pixel). Because of the variance, we do not know the true population mean ΔA_z(mean) of the normal distribution from which the ΔA_z(observed) was sampled. It can be estimated, however, that we have a less than 5% chance of observing this ΔA_z value if the population mean ΔA_z(mean) of the distribution is greater than [ΔA_z(observed) − $(-1.645) \cdot \sigma(\Delta A_z)$]. This estimated bound of mean ΔA_z is denoted as [max ΔA_z(mean)] and tabulated in Table 5.

Because an increasing sample size reduces only the variance while the population mean of the distribution of ΔA_z remains the same, the [max ΔA_z(mean)] estimated earlier for a finite sample size may also be considered to be the maximum mean ΔA_z for $N\rightarrow\infty$. Comparison of the values of [max ΔA_z(mean)], $\sigma(N\rightarrow\infty)$, and [min $\Delta A_z(N\rightarrow\infty)$] in Table 5 shows that the [max ΔA_z(mean)] is approximately equal to $\sigma(N\rightarrow\infty)$ and is thus smaller than [min $\Delta A_z(N\rightarrow\infty)$] for the 35- versus 70-$\mu$m and 35- versus 105-$\mu$m image pairs when the sample size approaches infinity. Therefore, the finite sample size in our

Table 2
Summary of A_z Values

Pixel Size (μm)	Reader 1	Reader 2	Reader 3	Reader 4	Reader 5	Reader 6	Reader 7	Average*
35	0.68	0.62	0.75	0.75	0.65	0.74	0.77	0.71
70	0.73	0.71	0.77	0.80	0.64	0.65	0.77	0.73
105	0.80	0.63	0.73	0.81	0.73	0.60	0.77	0.73
140	0.69	0.64	0.68	0.80	0.68	0.74	0.76	0.71

Note.—The standard deviations of the A_z values ranged from 0.05 to 0.07.
*A_z of average ROC curve, which was obtained by averaging the slope and intercept parameters of the individual ROC curves.



a. b.

Figure 6. Dependence of the A_z on pixel size for readers **(a)** 1, 4, and 5 and **(b)** 2, 3, 6, and 7.

Table 3
Comparison of 35-μm Images with 70, 105, and 140-μm Images

	All Readers		All Readers Except Reader 6	
Pixel Size (μm)	DBM Method	Paired *t* Test	DBM Method	Paired *t* Test
35 vs 70	.51	.51	.11	.10
35 vs 105	.65	.65	.12	.12
35 vs 140	.93	.91	.96	.96

Note.—Data are two-tailed *P* values.

current ROC study is not the main contributor to the lack of statistical significance in the difference for the 35- versus 70-μm and 35- versus 105-μm image pairs. The small difference in A_z relative to the large reader variance may be the main reason we did not observe a statistically significant advantage of the 35-μm pixel size over 70- or 105-μm pixel sizes in the characterization of malignant and benign microcalcifications.

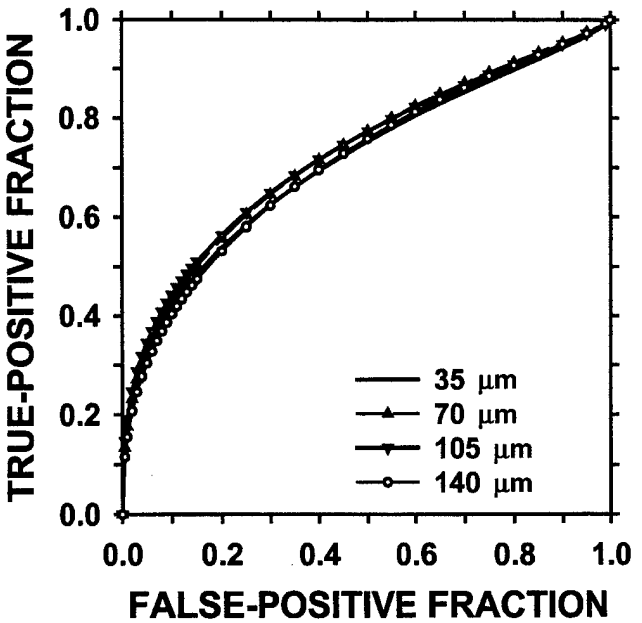


Figure 7. The average ROC curves for the four pixel sizes. Each curve was derived from the average slope and intercept parameters of the individual ROC curves from the seven radiologists.

Table 4
Percentage of Positive and Negative Cases that Received a Confidence Rating of 1

	35-μm Pixel Size		70-μm Pixel Size		105-μm Pixel Size		140-μm Pixel Size	
Reader	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
1	9.7	5.0	4.2	2.5	5.6	0.0	4.2	0.0
2	41.7	30.0	44.4	20.0	37.5	25.0	38.9	22.5
3	16.7	7.5	13.9	2.5	12.5	5.0	8.3	7.5
4	25.0	10.0	23.6	7.5	30.6	5.0	26.4	5.0
5	40.3	20.0	43.1	25.0	50.0	20.0	45.8	25.0
6	62.5	27.5	52.8	35.0	54.2	40.0	63.9	30.0
7	23.6	10.0	18.1	5.0	22.2	7.5	22.2	10.0

Table 5
Variance Components of the ROC Experiment

Modalities (μm)	σ^2_{mc}	σ^2_{mr}	σ^2_{ϵ}	$\sigma(\Delta A_z)$	ΔA_z(obs)	Max $\Delta A_z(m)$ at one-tailed, $P = .05$	$\sigma(N \to \infty)$	Min $\Delta A_z(N \to \infty)$ at one-tailed, $P = .05$
35 vs 70	−0.000009*	0.000867	0.000833	0.0216	−0.02	0.016	0.016	0.026
35 vs 105	−0.000014*	0.001803	0.000778	0.0266	−0.02	0.024	0.023	0.038
35 vs 140	0.000024	0.000488	0.000928	0.0213	−0.00	0.035	0.012	0.020
70 vs 105	0.000002	0.001213	0.000728	0.0236	−0.00	0.039	0.019	0.031
70 vs 140	0.000077	0.001195	0.000825	0.0270	0.02	0.064	0.018	0.030
105 vs 140	0.000031	0.001888	0.000763	0.0286	0.02	0.067	0.023	0.038

Note.—Data were estimated with the bootstrap method of Beiden et al (17). The total variance $\sigma^2(\Delta A_z)$ is computed from the variance components and Eq (21) of Beiden et al as $\sigma^2(\Delta A_z) = 2(\sigma^2_{mc} + \sigma^2_{mr}/R + \sigma^2_{\epsilon}/R)$, where R is the number of readers. Max $\Delta A_z(m)$ is the maximum mean difference in A_z between two modalities. $\sigma(N \to \infty) = (2\sigma^2_{mr}/R)^{1/2}$ is the standard deviation and Min $\Delta A_z(N \to \infty)$ is the minimum difference in A_z between two modalities that will allow rejection of the null hypothesis, A_z (small pixel) = A_z (large pixel) with $P < .05$ when the sample size N approaches infinity. The variance component σ^2_{mc} is negative in some cases due to the variance of the bootstrap estimation; the error bars tightly bracket the neighborhood of zero.

*Data are negative owing to the variance of the bootstrap estimation; their error bars tightly bracket the neighborhood of zero.

Another interesting observation can be made from the analysis of the variance components. In this classification task, the modality-by-case variance component σ^2_{mc} is consistently near zero for any of the paired comparisons. This means that even with an infinite number of readers, the variations in the two modalities will completely follow each other. It is still possible that the two modalities will have different mean performances, but cases that are more (or less) difficult with one modality will completely follow in the direction of cases that are more (or less) difficult with the other modality. This again seems to imply that the nature of the classification task is more dominant than the appearance of the image with each modality.

One aspect of the interobserver variabilities is demonstrated in Table 4, where the radiologists' decision thresholds for biopsy varied over a wide range. The large variation among the ROC curves in Figure 6 indicates that the variation among the radiologists' biopsy recommendation is not entirely caused by the use of a different decision threshold by each radiologist along similar ROC curves. This suggests that the estimation of the likelihood of malignancy of microcalcifications based on their mammographic features such as morphologic characteristics and spatial distribution pattern is very different among the radiologists. It may be noted, however, that the majority of the cases used in this ROC study had undergone biopsy so that easily distinguished benign cases had already been excluded from the case samples.

To investigate the intraobserver variabilities in the classification of microcalcifications, we repeated one reading session with three observers (readers 1–3). The distributions of the differences in the confidence ratings between the two readings of the same film of a cluster by

each radiologist are shown in Figure 8. The differences in the ratings range from −5 to +3 for reader 1, −5 to +6 for reader 2, and −3 to +4 for reader 3. This is consistent with the results of the variance analysis with the method of Beiden et al, where the reader variance was found to be an important component of the total variance for the classification of microcalcifications.

We also attempted to analyze the correlation of the estimated likelihood of malignancy when the same images were read by different radiologists. The scatter plots of the malignancy ratings by every two radiologists (not shown) were, in general, spread over wide ranges without obvious correlation. The histograms of the difference in the malignancy ratings for the same cluster between two radiologists were similar to those of the intraobserver variability shown in Figure 8, with ranges as wide as −6 to +6. There were some trends that some radiologists (eg, reader 1) tended to have higher likelihood of malignancy estimates for most clusters than did other radiologists, and some radiologists (eg, reader 6) tended to have lower suspicion for malignancy than did the others. These trends are consistent with the lower biopsy threshold of reader 1 and the higher biopsy threshold of reader 6 (Table 4).

We investigated whether the intraobserver variability in the malignancy ratings depended on the perceived subtlety of the microcalcification cluster. Because reader 3 demonstrated the smallest range of variability in the malignancy ratings among the three radiologists with whom we repeated the experiment, we plotted the relationship of the difference in the malignancy ratings between the two readings of the same cluster against the subtlety rating of the cluster for reader 3, as shown in Figure 9. There was no obvious correlation between the variability in the malignancy ratings and the perceived subtlety of the clusters.

The large inter- and intraobserver variabilities in the malignancy ratings may be a result of the fact that the radiologists usually do not have to estimate specifically the likelihood of malignancy of the clusters when they read mammograms in clinical practice. However, because their decision threshold for biopsy recommendation also varied over a wide range, as discussed above, the variabilities were not simply caused by their unfamiliarity in the estimation of the likelihood of malignancy. The variabilities may again reflect the low specificity of the image features of the microcalcifications. As can be seen from the examples in Figure 3, the appearance of a cluster of benign microcalcifications from sclerosing adenosis can be very similar to that of a malignant cluster from intraductal carcinoma.

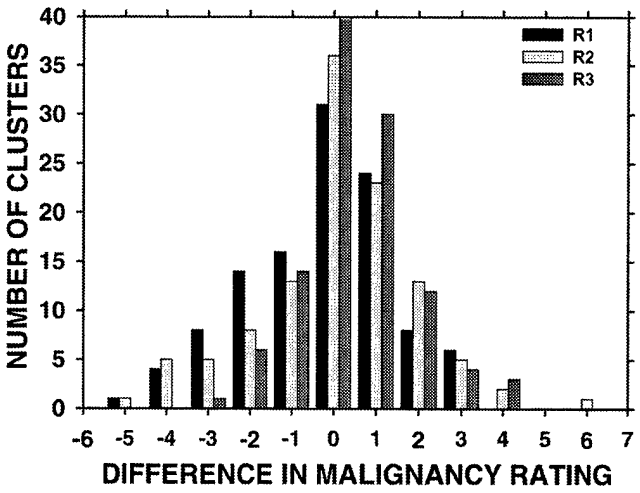


Figure 8. The distributions of the differences in the confidence ratings between the two readings of the same film of a cluster by the same radiologist for readers 1–3.

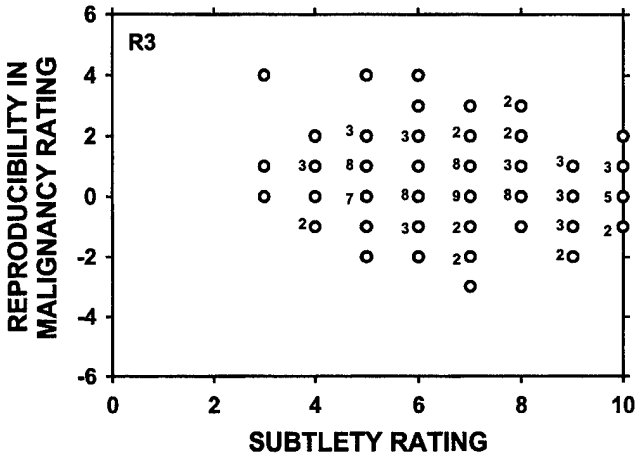


Figure 9. Scatter plot shows the relationship between the differences in confidence ratings between the two readings of the same cluster and the subtlety ratings of the cluster, as rated by reader 3. The number next to a data point indicates the number of cases that overlap at the same point. Data points without a number indicate that there is only one case at that point.

The dependence of classification accuracy on pixel size may be further weakened when other patient information is available for making diagnostic decisions. In clinical practice, the decision for biopsy is not dependent on the mammographic appearance alone. When the morphologic information is nonspecific, other patient information (eg, age, family history, and personal history) becomes important for estimating the likelihood of breast cancer. Because our goal was to evaluate whether the classification accuracy of microcalcifications depended on the pixel size of the digitized images, we did not provide such patient

information to the observers. Our results indicate that the mammographic information that a radiologist assesses from the displayed images, such as the morphologic characteristics and spatial distribution pattern of the microcalcifications, does not have a strong dependence on pixel size in the range studied.

It may be noted that in our current ROC study we concentrated on the effect of pixel size on the classification of malignant and benign microcalcifications according to their mammographic features. We previously conducted an ROC study (9) to compare the detectability of subtle microcalcifications on original screen-film mammograms with that on mammograms digitized at a pixel size of 100 μm by using an optical drum scanner. We found that the detection accuracy for the subtle microcalcifications decreased when radiologists read the 100-μm pixel size digitized images. Results of another previous study (10), in which we investigated the detection of microcalcifications by a computer program, also indicated a reduction in detectability when the digitization pixel size increased from 35 to 140 μm. The results from these experiments indicate that spatial resolution may be more important for the detection than for the classification of microcalcifications in mammographic imaging.

In clinical practice, an important technique used by radiologists to estimate the likelihood of malignancy of a microcalcification cluster is to evaluate its interval change between examinations. The number of microcalcifications in a cluster is an important feature for characterizing changes. High-quality mammograms that can provide sensitive detection of new, subtle microcalcifications are crucial for such a task. The results of our previous studies (9,10) indicate that the spatial resolution of mammographic images will affect the detectability of subtle microcalcifications. The pixel size of digital mammograms may, therefore, affect the evaluation of interval changes, although the effect will be reduced with the use of magnification views. Because the radiologists in our current study were not provided with images from previous examinations for comparison, the effects of pixel size on the detection of interval change will warrant further investigation.

Another possible reason that the images with a 35-μm pixel size did not provide better classification accuracy for malignant and benign microcalcifications than did images with 70- or 105-μm pixel sizes, as observed in this study, is the higher noise level in the digitized images at this small pixel size. A higher noise level will reduce the signal-to-noise ratio of the image and may interfere with the perception of image features. It is possible that if the radiation dose to the patient is unlimited, a digital mammography system with a smaller pixel size can provide better classification. In the current study, we investigated the dependence of classification accuracy on pixel size under the constraint of equal radiation dose. The trade-off between image quality and radiation dose and the acceptability of higher-dose techniques are beyond the scope of this study. Furthermore, because digitized mammograms and mammograms acquired with digital detectors have different noise, contrast sensitivity, and resolution properties, further investigations are needed to determine whether a similar trend holds for mammograms acquired with different types of digital detectors.

In conclusion, we performed an ROC study to investigate the effects of pixel size on the classification of malignant and benign microcalcifications on digitized mammograms. Our results indicate that the differences in the A_z between pairs of pixel sizes ranging from 35 to 140 μm do not achieve statistical significance. The pixel sizes in this range therefore do not have a strong effect on radiologists' accuracy in the characterization of microcalcifications. The low specificity of the image features of microcalcifications and the large interobserver and intraobserver variabilities may have prevented small advantages in image resolution from being observed.

ACKNOWLEDGMENTS

The authors are grateful to Sergey V. Beiden, PhD, for analysis of the variance components, Charles E. Metz, PhD, for the LABMRMC program, and Robert F. Wagner, PhD, and Charles E. Metz, PhD, for helpful discussion on statistical analysis.

REFERENCES

1. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics, 1998. CA Cancer J Clin 1998; 48:6–29.
2. Byrne C, Smart CR, Cherk C, Hartmann WH. Survival advantage differences by age: evaluation of the extended follow-up of the Breast Cancer Detection Demonstration Project. Cancer 1994; 74:301–310.
3. Feig SA, Hendrick RE. Risk, benefit, and controversies in mammographic screening. In: Haus AG, Yaffe MJ, eds. Syllabus: a categorical course in physics—technical aspects of breast imaging. Oak Brook, Ill: Radiological Society of North America, 1993; 119–135.
4. Tabar L, Dean PB. Teaching atlas of mammography. New York, NY: Thieme, 1985.
5. Wolfe JN. Analysis of 462 breast carcinomas. AJR Am J Roentgenol 1974; 121:846–853.
6. Murphy WA, DeSchryver-Kecskemeti K. Isolated clustered microcalcification in the breast: radiologic-pathologic correlation. Radiology 1978; 127:335–341.

7. Millis RR, Davis R, Stacey AJ. The detection and significance of calcifications in the breast: a radiological and pathological study. Br J Radiol 1976; 49:12–26.
8. Sickles EA. Mammographic features of 300 consecutive nonpalpable breast cancers. AJR Am J Roentgenol 1986; 146:661–663.
9. Chan HP, Vyborny CJ, MacMahon H, Metz CE, Doi K, Sickles EA. Digital mammography: ROC studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications. Invest Radiol 1987; 22:581–589.
10. Chan HP, Niklason LT, Ikeda DM, Lam KL, Adler DD. Digitization requirements in mammography: effects on computer-aided detection of microcalcifications. Med Phys 1994; 21:1203–1211.
11. Karssemeijer N, Frieling JTM, Hendriks JHCL. Spatial resolution in digital mammography. Invest Radiol 1993; 28:413–419.
12. Kallergi M, Clarke LP, Qian W, et al. Interpretation of calcifications in screen-film, digitized, and wavelet-enhanced monitor-displayed mammograms: a receiver operating characteristic study. Acad Radiol 1996; 3:285–293.
13. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
14. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. Radiology 1991; 179:463–468.
15. Sickles EA. Nonpalpable, circumscribed, noncalcified solid breast masses: likelihood of malignancy based on lesion size and age of patient. Radiology 1994; 192:439–442.
16. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol 1992; 27:723–731.
17. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative methodology for random-effects ROC analysis. Acad Radiol 2000; 7:341–349.

Improvement of mammographic mass characterization using spiculation measures and morphological features

Berkman Sahiner,[a] Heang-Ping Chan, Nicholas Petrick, Mark A. Helvie, and Lubomir M. Hadjiiski
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 4 December 2000; accepted for publication 3 May 2001)

We are developing new computer vision techniques for characterization of breast masses on mammograms. We had previously developed a characterization method based on texture features. The goal of the present work was to improve our characterization method by making use of morphological features. Toward this goal, we have developed a fully automated, three-stage segmentation method that includes clustering, active contour, and spiculation detection stages. After segmentation, morphological features describing the shape of the mass were extracted. Texture features were also extracted from a band of pixels surrounding the mass. Stepwise feature selection and linear discriminant analysis were employed in the morphological, texture, and combined feature spaces for classifier design. The classification accuracy was evaluated using the area A_z under the receiver operating characteristic curve. A data set containing 249 films from 102 patients was used. When the leave-one-case-out method was applied to partition the data set into trainers and testers, the average test A_z for the task of classifying the mass on a single mammographic view was 0.83 ± 0.02, 0.84 ± 0.02, and 0.87 ± 0.02 in the morphological, texture, and combined feature spaces, respectively. The improvement obtained by supplementing texture features with morphological features in classification was statistically significant ($p = 0.04$). For classifying a mass as malignant or benign, we combined the leave-one-case-out discriminant scores from different views of a mass to obtain a summary score. In this task, the test A_z value using the combined feature space was 0.91 ± 0.02. Our results indicate that combining texture features with morphological features extracted from automatically segmented mass boundaries will be an effective approach for computer-aided characterization of mammographic masses. [DOI: 10.1118/1.1381548]

Key words: computer-aided diagnosis, mammography, breast mass characterization, segmentation, morphological features

I. INTRODUCTION

Mammography is currently the only proven and cost-effective method to detect early breast cancer. Masses are important indicators of malignancy on mammograms. However, only a small percentage of masses found on mammograms are malignant. Many benign conditions, such as cysts and fibroadenomas are detected as breast masses. Some benign masses may look suspicious enough for the radiologist to recommend biopsy. In three studies, it was found that only 20%–30% of mammographically suspicious nonpalpable breast masses that underwent biopsy were malignant.[1–3] In order to reduce costs and patient discomfort, it is important to reduce the number of benign biopsies without missing any malignant masses. Computer-aided diagnosis has the potential to assist the radiologists in the characterization of mammographic masses.[4]

In recent years, many researchers have investigated the use of computer-extracted image features for classification of breast masses as malignant or benign. The features were extracted from the gray-level and morphological characteristics of the lesion. Kilday *et al.*[5] extracted mass shapes using interactive gray-level thresholding, and classified them into cancer, cyst, and fibroadenoma categories using morphological features and patient age. Pohlman *et al.*[6] segmented masses using an adaptive region growing algorithm, whose parameters were interactively adjusted. After mass segmentation, features related to tumor shape and boundary roughness were automatically extracted and used for the classification of the lesions. They found that their tumor boundary roughness feature provided slightly inferior classification accuracy compared to two experienced radiologists who specialized in mammography. Rangayyan *et al.*[7] used a measure of the diffusion of a mass into the surrounding mammogram termed edge acutance, as well as a number of shape factors, including Fourier descriptors, moments, and compactness, to classify masses. They found the edge acutance measure to be superior to the other features extracted from the mass shape. Using the acutance measure alone, they were able to correctly classify 93% of masses in a database of 54 cases. Viton *et al.*[8] characterized the degree of spiculation and the presence of fuzzy areas in the region surrounding a mass by means of polar and pseudopolar representations of this region. Huo *et al.*[9] extracted features related to the margin and the density of the masses for classification. They designed and tested a two-stage hybrid classifier consisting of a rule-based stage and an artificial neural network stage on a data

set of 95 mammograms. The hybrid classifier achieved an area under the receiver operating characteristic (ROC) curve of 0.94 for their data set. Sahiner *et al.* and Chan *et al.* used texture features extracted from transformed images for characterization of breast masses,[10] and investigated the effect of their CAD method on radiologists' rating of breast masses.[4] They showed that their CAD method could significantly improve radiologists' accuracy in characterization of masses, and thereby might reduce unnecessary biopsies.

A second class of techniques for computer aided characterization of breast lesions use the computer to combine mammographic features extracted by a radiologist into a malignancy rating. Getty *et al.* designed a classifier based on 12 mammographic features extracted by radiologists, and showed that the classifier could substantially increase the radiologists' diagnostic accuracy.[11] Lo *et al.* and Baker *et al.* designed a neural network classifier based on BI-RADS features of the American College of Radiology, and the personal and family history of the patient.[12–14] The neural network classifier had significantly higher specificity at high sensitivity levels compared to radiologists.[14]

In the clinical evaluation of a mammographic mass, its shape and margin characteristics are very important.[15] We previously introduced a rubber-band straightening transform to analyze the margin characteristics of a mass.[10] In the present study, our aim is to include features related to the shape of the mass to improve the characterization accuracy. In order to obtain an accurate delineation of mass boundaries, we have developed a fully automated three-stage segmentation method. The first stage of our segmentation method is based on a clustering technique that we previously investigated. Clustering is used to find the general outline of the mass shape. This general outline is refined using an active contour method in the second stage. In the third stage, spiculations are detected and segmented based on image gradient directions. After segmentation, morphological features are extracted from the mass shape, and are combined with the texture features that we have previously utilized for characterization of breast masses.

II. METHODS

A. Data set

The mammograms used in this study were randomly selected from the files of patients in the Radiology Department at the University of Michigan who had undergone biopsy. All mammograms were acquired with dedicated mammographic systems. The criteria for inclusion of a mammogram in the data set were that the mammogram contained a biopsy-proven mass, and that approximately equal numbers of malignant and benign masses were present in the data set.

Our data set consisted of 249 mammograms from 102 patients. The mammograms contained a total of 122 benign and 127 malignant masses. The true pathology of the masses was determined by biopsy and histologic analysis. Six of the benign masses, and 63 of the malignant masses were characterized as spiculated by a radiologist experienced in mammographic interpretation. Out of the 249 mammograms, 223

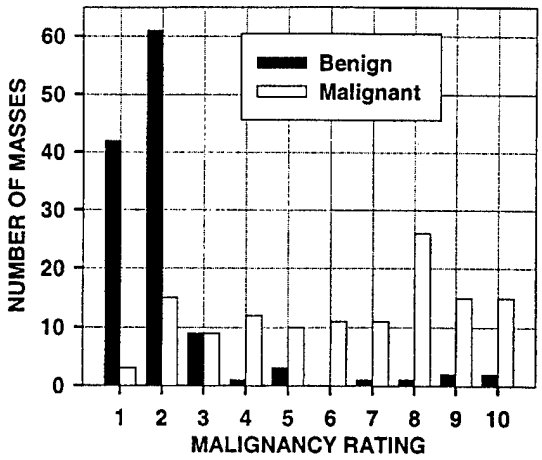


FIG. 1. The distribution of the malignancy rating of the masses in our data set, by an experienced radiologist: (1) very likely benign, (10) very likely malignant.

were acquired six months or less before biopsy, and 26 were acquired more than six months before biopsy. The probability of malignancy of the biopsied mass on each mammogram was ranked by a Mammography Quality Standards Act (MQSA) approved radiologist on a scale of 1 (most benign mammographic appearance) to 10 (most malignant mammographic appearance). The distribution of the malignancy ranking of the masses on each view is shown in Fig. 1. Note that the malignant and benign masses overlap over the entire range of suspicion for malignancy, indicating that the malignant or benign features of these masses could not be easily distinguished by radiologists. This is consistent with the fact that all these masses had undergone biopsy. The size of the masses in our data set ranged from 5 to 29 mm (mean size =12.5 mm). The distribution of the size for malignant and benign masses is shown in Fig. 2. It is observed that the distribution of the size for malignant masses is similar to that for benign masses.

The mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel size of 100 μm×100 μm and

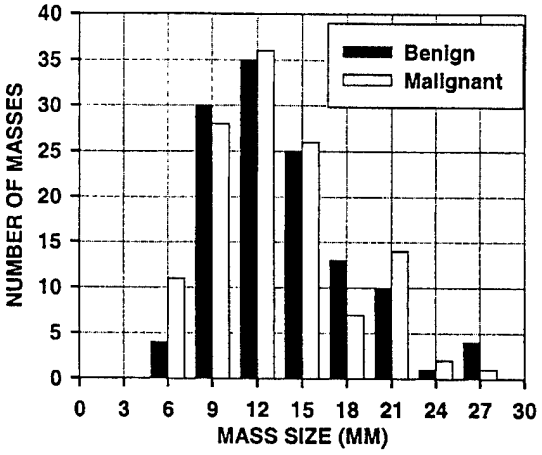


FIG. 2. The distribution of the mass size for the 249 masses in our data set. Mass sizes were measured as the longest dimension of the mass by an experienced radiologist.

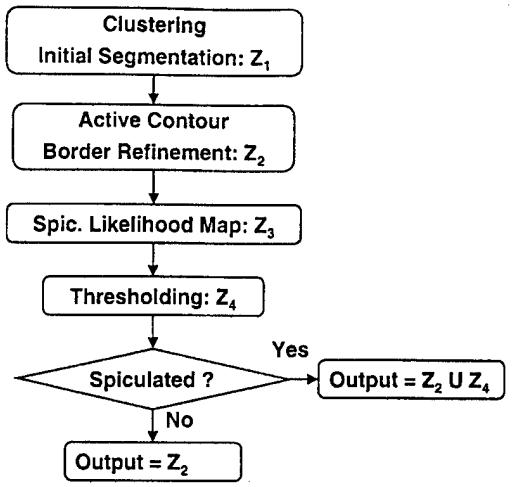


FIG. 3. The block diagram for the mass segmentation algorithm. All images Z_k, for $k \neq 3$, are binary images, with a nonzero value indicating an object pixel.

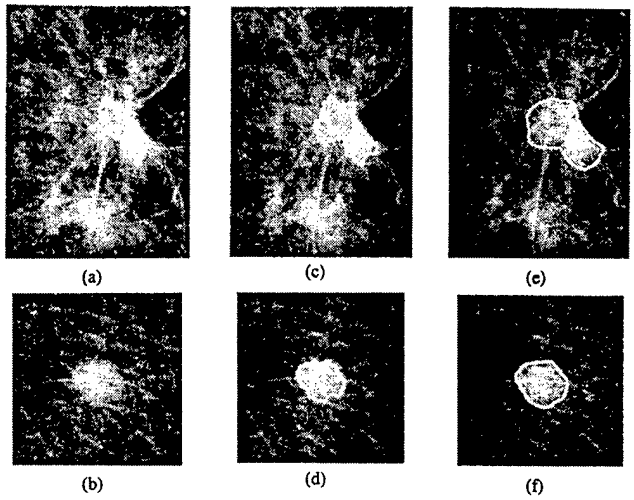


FIG. 4. The mass ROI, the initial contour, and the final contour of the active contour model for a spiculated mass [(a), (c), and (e)] and a nonspiculated mass [(b), (d), and (f)].

4096 gray levels. The digitizer was calibrated so that gray level values were linearly proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of 0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually, with the OD range extending to 3.5. The pixel values were linearly converted before they were stored on the computer so that a high pixel value represented a low optical density.

The location of the biopsied mass was identified by the radiologist, and a region of interest (ROI) containing the mass was extracted for computerized analysis. The size of the ROI was chosen such that the radiologist-marked lesion and a band of about 50-pixel-wide surrounding background were included in the ROI.

Before any processing, the ROIs were first processed with a background correction algorithm. The goal of background correction is to reduce the nonuniform background caused by the overlapping breast structures and the location of the lesion on the mammogram. The nonuniform background is not related to mass malignancy, but may affect the segmentation and feature extraction results used in our computerized analysis. Details and examples of our background correction technique can be found in the literature.[16,17]

B. Mass segmentation

We used a fully automated segmentation method to extract the mass shape. The block diagram for our mass segmentation algorithm is shown in Fig. 3, and the individual steps of the segmentation algorithm are explained in the following.

1. Initial mass segmentation

The mass segmentation method employed in this study started with the initial detection of a mass shape within a ROI using a pixel-by-pixel K-means clustering algorithm, which was discussed in detail in the literature.[18,19] The parameters of the segmentation algorithm were chosen so that the segmented region was slightly smaller than the apparent size of the mass. This choice prevented most of the masses from merging into neighboring objects. After clustering, one to several objects would be segmented in the ROI. If more than one object was segmented, the largest connected object was selected. The selected object was then filled, grown in a local neighborhood, and eroded and dilated with morphological operators. In the resulting binary image, a nonzero value indicated an object pixel, and zero value indicated a background pixel. The implementation details of these steps have been described in the literature.[10] Figures 4(a)–4(d) show examples of a spiculated mass and a nonspiculated mass and the results of the first stage segmentation.

2. Active contour segmentation

Although initial mass segmentation resulted in reasonable mass shapes for most of the masses, further refinement was necessary before detection and segmentation of the spiculations. We used an active contour model for mass shape refinement.

An active contour is a deformable continuous curve, whose shape is controlled by internal forces (the model, or *a priori* knowledge about the object to be segmented) and external forces (the image).[20] The internal forces impose a smoothness constraint on the contour, and the external forces push the contour toward salient image features, such as edges. To solve a segmentation problem, an initial boundary is iteratively deformed so that the energy due to internal and external forces is minimized along the contour. The energy terms used in our implementation are described in the literature.[21] We used the shape segmented by our first stage segmentation method as the initial boundary. To minimize the contour energy, we used an iterative algorithm proposed by Williams and Shah.[22] The details of our active contour model have been described elsewhere.[23] Figures 4(c)–4(f) show the initial and final contours of the model for a spiculated mass and a nonspiculated mass, respectively. A binary image, denoted by Z_2 in the schematic shown in Fig. 3, is

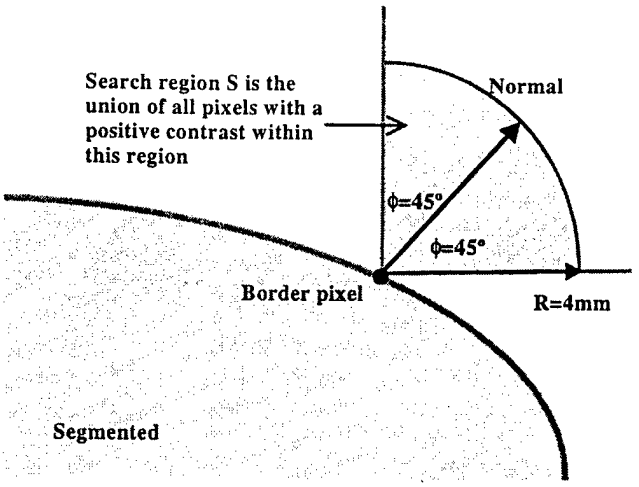


FIG. 5. The definition of the search region for a given border pixel.

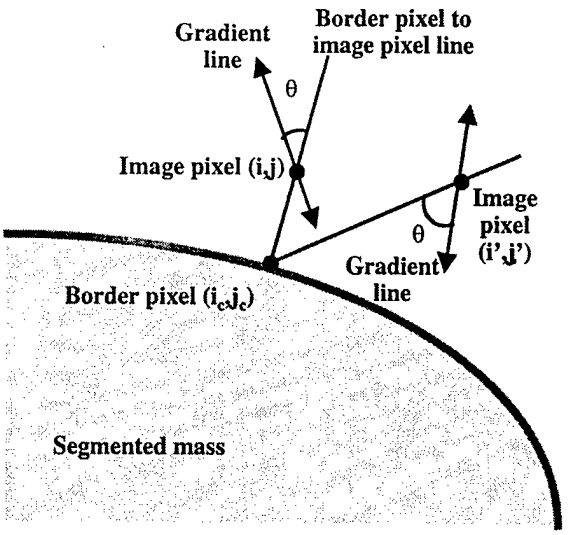


FIG. 6. The definition of the angular difference θ.

produced by filling the interior of the resulting contour, such that any pixel within the object has a pixel value of 1, and any background pixel has a pixel value of 0.

3. Segmentation of spiculations

Spiculations on mammograms appear as linear structures with a positive image contrast, and they usually lie in a radial direction to the mass. As a result of their linearity, the gradient directions at image pixels on or close to the spiculation are more or less in the same orientation relative to that of the spiculation. Karssemeijer *et al.* have used this property for detecting spiculated lesions on mammograms.[24] In this study, we developed a method for determining whether a pixel (i_c, j_c) on the mass contour lies on the path of a spiculation, and to segment the spiculation if it does.

For a pixel (i_c, j_c) on the mass boundary, a search region $S(i_c, j_c)$ is defined as the set of all image pixels that (i) lie outside the mass; (ii) have a positive contrast; (iii) are at a distance less than 4 mm from (i_c, j_c); and (iv) are within $\pm\pi/4$ of the normal to the mass contour at (i_c, j_c) (Fig. 5). At each image pixel (i,j) in $S(i_c, j_c)$, the obtuse angle θ between two lines is computed, where the first line is defined by the gradient direction at (i,j), and the second line joins the pixel (i,j) to the mass boundary pixel (i_c, j_c) (Fig. 6). We have used a method based on convolution with Gaussian derivatives[25] for computing the gradients. The spiculation measure $x(i_c, j_c)$ at a mass boundary pixel (i_c, j_c) is defined as the average value of θ in the search region $S(i_c, j_c)$. If the pixel (i_c, j_c) lies on the path of a spiculation, then θ will be close to $\pi/2$ whenever the image pixel (i,j) is on the spiculation, and hence the mean of the spiculation measure will be high.

For the segmentation task, we computed $x(i_c, j_c)$ for a sequence of 30 contours. The first contour in the sequence is that provided by the active contour model. The following contours in the sequence are obtained by expanding the previous contour by one pixel at a time, so that x is computed in a 30-pixel-wide band around the mass. The resulting image in the 30-pixel-wide band around is referred to as the spiculation likelihood map, and is denoted by Z_3 in Fig. 3. Figure 7 shows the spiculation likelihood map for the two masses used in Fig. 4. The spiculation likelihood map Z_3 is used for both detecting whether a mass is spiculated, and for segmenting the spiculations. To detect whether a mass is spiculated, a binary image Z_4 is produced by thresholding Z_3, at a threshold T. After initial experimentation, the value of T was chosen to be 0.85. This threshold was kept constant in the segmentation algorithm for all images used in the study.

After thresholding, all connected objects in Z_4 are detected. The number of the objects is used as an estimate of the number of possible spiculations. The ratio of the total area of the objects in Z_4 to the mass area is used as an indication of the relative size of the spiculations. The product of the two features above (number of objects and the size ratio) is used as a *spiculation detection variable* to classify the mass as spiculated or nonspiculated. The choice of the threshold for this classification is discussed in Sec. II D. If the mass is classified as spiculated, then the algorithm combines the binary image that represents the mass outline detected by the active contour model (Z_2) and the binary image

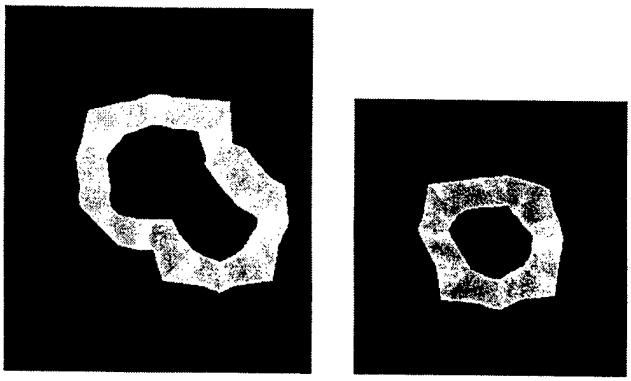
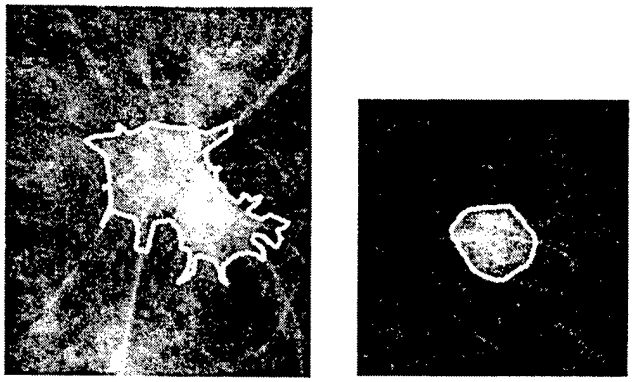


FIG. 7. The spiculation likelihood maps for the spiculated and the nonspiculated masses shown in Fig. 4: (a) spiculated, (b) nonspiculated.



(a) (b)

FIG. 8. The result of the final segmentation for the spiculated and the nonspiculated masses shown in Figs. 4 and 7: (a) spiculated, (b) nonspiculated.

that represents the result of thresholding (Z_4) to segment the spiculations (Fig. 3). If the mass is classified as nonspiculated, then the output of the segmentation is Z_2. Figure 8 shows the result of spiculation detection and segmentation for the masses used in Figs. 4 and 7.

C. Feature extraction

1. Extraction of morphological features

Malignant masses tend to have more irregular contours than benign masses. In addition, spiculation is a strong indication for malignancy. Therefore, features related to the segmented mass shape are expected to yield useful information for characterization of breast masses. In this study, thirteen morphological features were extracted from the final mass outline. A list of these thirteen features, as well as their accuracy in classifying each mass in our data set as malignant or benign, are shown in Table I. In this section, we describe these morphological features. The classification accuracy is discussed in Sec. IV.

The first five morphological features listed in Table I are based on the normalized radial length (NRL), defined as the Euclidean distance from the object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object.[5] In our previous studies, we found that NRL mean, standard deviation, entropy, area ratio, and zero crossing count were useful for discriminating between objects containing masses and normal tissue.[26]

TABLE I. The list of the morphological features used in this study, and the area A_z under the ROC curve when each feature is used alone for classification.

Morphological feature name	Classification accuracy A_z
Fourier descriptor	0.82
Convexity	0.79
Rectangularity	0.75
Perimeter	0.75
NRL mean	0.72
Contrast	0.71
NRL entropy	0.69
Circularity	0.67
NRL area ratio	0.66
NRL standard deviation	0.65
NRL zero crossing count	0.64
Perimeter-to-area ratio	0.63
Area	0.60

The sixth feature, convexity, is defined as the ratio of the area of the segmented object to the area of the smallest convex shape that contains the object. If the object is convex, as is the case with many benign masses, then this feature will attain its maximum value of unity. If the object shape is highly nonconvex, as is the case with many spiculated or malignant masses, then the value of this feature will be small.

The seventh feature, Fourier descriptor (FD), is based on the Fourier transform of the object boundary sequence. To compute the Fourier transform of the object boundary sequence, the x and y coordinates of each border pixel m is represented as a complex number, $z(m)=x(m)+jy(m)$, where $0m<N$, and $z(m)$ is a periodic sequence with period N. Let $c(k)$ denote the Fourier coefficients of the periodic sequence $z(m)$, and let $d(k)$ be a periodic sequence with period N, defined in the interval $0k<N$ as

$$d(k)=\begin{cases}0 & k=0\\ |c(k)/c(1)| & k\neq 0.\end{cases} \tag{1}$$

It can be shown that $d(k)$ is independent of rotation, translation, and scaling of the object, and the choice of the initial point $z(0)$ on the object contour sequence.[27] Objects with irregular contours have more high-frequency components than those with smooth contours. The following summary Fourier descriptor measure[28] which emphasizes low-frequency components of $d(k)$ is therefore useful in discriminating between shapes with smooth and irregular contours

$$\mathrm{FD}=\frac{\sum_{k=-N/2+1,k\neq 0}^{N/2} d(k)/|k|}{\sum_{k=-N/2+1,k\neq 0}^{N/2} d(k)}. \tag{2}$$

For computational efficiency, all contours were interpolated to a large integral power of 2, (2^{12}) before the computation of the Fourier series.

The remaining six features were also shown to be useful in discriminating between objects containing masses and normal tissue.[26] These features include the perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast of the object. The definition of these features can be found in the literature.[26]

2. Extraction of texture features

The texture of the region surrounding the mass can yield important features for its classification. Since possible spiculations and the gradient of the opacity caused by the mass are approximately radially oriented, the texture of the region surrounding a mass is expected to have a radial dependence. However, most texture extraction methods are designed for texture orientations in a uniform direction (horizontal, verti-

TABLE II. The list of the texture features used in this study, and the area A_z under the ROC curve when each feature is used alone for classification. For each measure, the range of A_z values for different pixel-pair distances and directions is shown.

Spatial gray-level dependence (SGLD) feature measure	Classification accuracy A_z	Run-length statistics (RLS) feature measure	Classification accuracy A_z
Difference average	0.52–0.66	Long runs emphasis	0.63–0.66
Difference entropy	0.53–0.66	Run percentage	0.59–0.65
Inverse difference moment	0.50–0.66	Gray level nonuniformity	0.59–0.62
Difference variance	0.52–0.65	Run length nonuniformity	0.55–0.57
Inertia	0.53–0.65	Short runs emphasis	0.50–0.56
Correlation	0.50–0.61		
Inf. measure of correlation 1	0.50–0.61		
Inf. measure of correlation 2	0.50–0.59		
Energy	0.54–0.59		
Entropy	0.54–0.58		
Sum variance	0.52–0.58		
Sum entropy	0.51–0.57		
Sum average	0.55–0.56		

cal, or at a certain angle between these two directions). To be able to extract meaningful texture features from the region surrounding a mass, we have designed a rubber band straightening transform (RBST) that maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region).[10,29,30] In the transformed image, the border of the mass is expected to appear approximately as a horizontal edge, and spiculations are expected to appear approximately as vertical lines.

The mass outline produced by the first stage segmentation discussed previously is used for defining the RBST image. The mass object produced by this stage is usually slightly smaller than what can be visually discerned on the mammogram. Thus, a thin border region along the mass margin is included in the RBST image. Important texture and gradient information at the mass margin is therefore included in the analysis of the region surrounding the mass. A 40-pixel-wide region, corresponding to a 4 mm band is used to determine the RBST image.

The texture features extracted from the RBST images include 13 texture measures, each calculated at 4 directions and 10 distances, from the spatial gray-level dependence (SGLD) matrices, and 5 run-length statistics (RLS) measures, each calculated at four directions, as described in our previous work.[10] A list of the SGLD and RLS texture measures is shown in Table II. Also shown in Table II are the classification accuracies when each measure is used alone to distinguish between malignant and benign ROIs. For conciseness, the range of classification accuracy (over four directions and ten distances for SGLD measures, and over four directions for RLS measures) of each texture measure is shown. The definition of these features[31,32] and the parameters used in this study can be found in the literature.[10]

D. Classification

The classifier in this study was designed to classify the masses on each available view. The same mass imaged on the CC and MLO views, and any additional views received different classification scores for each view. To assess the classifier accuracy, we considered both film-based and case-based methods. In the film-based method, the purpose was to classify the mass on each view as malignant or benign. In the case-based method, the purpose was to classify each mass as malignant or benign, using the information from all available views. To merge the information from different views of a lesion, we considered two methods. In the first method, the scores from different views were averaged. In the second method, the maximum malignancy score among all views was used as the score of the mass. The second method corresponds to calling a mass malignant if it appears to be malignant on any view, whereas the first method gives equal weight to each view to predict malignancy.

Stepwise feature selection and linear discriminant analysis were used for classifier design, and an N-fold cross-validation resampling scheme was used for partitioning the data into design and test sets. In a first set of experiments, we used tenfold cross validation. The data set was partitioned into ten random partitions such that all mammograms from one patient were grouped into the same partition. Nine of the partitions were used for feature selection and classifier training, and the remaining partition was used for testing. The purpose of grouping all mammograms of one patient into the same partition was to ensure that the test data were independent from training. Without this type of partitioning, one mammogram from a patient may be used for training a classifier that will be tested on another mammogram of the same patient, which may bias the test results because the training and test sets may not be completely independent. The test partition was rotated in a round-robin manner so that all partitions served as a test partition once and only once. The discriminant scores were analyzed using ROC methodology, using the LABROC program of Metz *et al.*[33] For each test partition, the classification accuracy was evaluated as the area A_z under the ROC curve. A mean A_z value for the data set was obtained by averaging these ten A_z values. In a second set of experiments, we used a leave-one-case-out method for data partitioning. This method is similar to ten-fold cross validation discussed previously, with the differences that, in

the leave-one-case-out method, each partition consisted of films from one and only one patient, and that the scores from all ROIs were accumulated for the ROC analysis. Since there were 102 patients, this corresponded to 102-fold cross validation. The statistical significance of the difference between ROC curves obtained with classifiers using different feature spaces (texture, morphological, or combined) was tested using the CLABROC program of Metz *et al.*[34]

Classifier training consisted of three stages, and was based on the training set alone for all of these three stages. The first stage was related to mass segmentation. As discussed in Sec. II B, the decision to classify a mass as spiculated or nonspiculated was based on thresholding a spiculation detection variable obtained from the spiculation likelihood map. The value of this threshold was determined from the training set such that the sum of correct decision percentages for the spiculated and nonspiculated masses was maximized for the training set. Classification of a mass as spiculated or nonspiculated determined if the spiculation segmentation step would be applied to the mass (see Fig. 3). This affected the morphological features extracted and selected in the second stage of classifier training. The second stage of the training involved stepwise feature selection,[35,36] which has been used for classifier design in many of our CAD applications.[10,17,37,38] Stepwise feature selection iteratively enters features into or removes features from the group of selected features based on a feature selection criterion. In this study, the feature selection criterion was based on the Wilks' lambda,[39] obtained using the trainers alone. The number of features in stepwise feature selection was controlled by the F-to-enter and F-to-remove thresholds, which were evaluated over a range from 5.0 to 2.0. In the third stage, the coefficients of the linear classifier were determined based on the training set. By making these three decisions independent of the test set, we aimed at improving the generalizability of our classification results to unknown cases in the patient population.

III. RESULTS

Figure 9 shows the distribution of the detection variable used for the classification of a mass as spiculated or nonspiculated. It is observed that by properly choosing the threshold, more than 30% (60/180) of the nonspiculated masses can be correctly identified without misclassifying any spiculated masses. At the selected threshold for the spiculation detection variable (see the earlier paragraph) 77% (53/69) of the spiculated masses and 78% (140/180) of the nonspiculated masses were correctly identified. Since there are six spiculated but benign masses in our data set, we did not use this variable for the classification of the masses as malignant or benign.

For both the tenfold cross validation and leave-one-case-out data partitioning methods, we investigated the classification of the masses as malignant or benign in the morphological feature space alone, texture feature space alone, and the combined morphological and texture feature space.

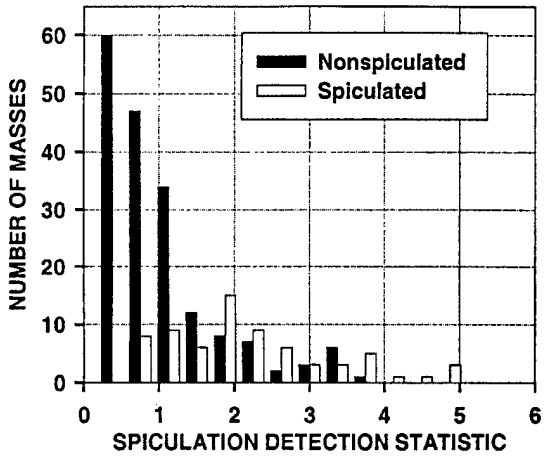


FIG. 9. The distribution of the spiculation detection variable for the spiculated and the nonspiculated masses.

A. Tenfold cross validation

The average number of selected features was 2, 10, and 14 in the morphological, texture, and combined feature spaces. The resulting A_z values for each of the ten partitions are shown in Table III. It is observed that combining the morphological and texture feature spaces improves the classification accuracy. The average A_z value for the ten partitions in this study was 0.85 for either the texture or the morphological features used alone. Using the combined feature space, the average test A_z value for the ten partitions reached 0.89.

B. Leave-one-case-out

The average number of selected features was 4, 8, and 10 in the morphological, texture, and combined feature spaces. The resulting A_z values were 0.84 ± 0.02, 0.83 ± 0.02, and 0.87 ± 0.02 in the morphological, texture, and combined feature spaces, respectively. The ROC curves for classification in these three feature spaces is shown in Fig. 10. For classification in the combined feature space ($A_z=0.87\pm0.02$), the distribution of the classifier scores for the 249 masses is shown in Fig. 11. This distribution represents film-based classification results, in the sense that the mass on each film

TABLE III. The test A_z values for each partition using linear discriminant analysis with morphological, texture, and combined feature spaces.

Partition number	Morphological feature space	Texture feature space	Combined feature space
1	0.90±0.06	0.92±0.06	0.92±0.07
2	0.92±0.06	0.98±0.03	1.000 000
3	0.83±0.10	0.93±0.06	0.94±0.05
4	0.80±0.08	0.83±0.08	0.86±0.08
5	0.94±0.05	0.80±0.16	0.92±0.07
6	0.82±0.08	0.66±0.12	0.85±0.08
7	1.000 000	1.000 000	0.96±0.04
8	0.77±0.10	0.71±0.10	0.71±0.11
9	0.64±0.11	0.73±0.10	0.74±0.10
10	0.93±0.05	0.91±0.06	0.98±0.03
Average	0.85	0.85	0.89

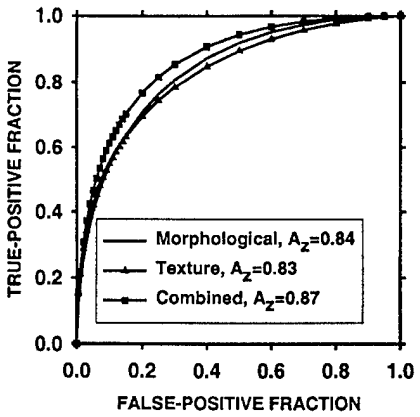


FIG. 10. ROC curves for classification of masses in the morphological, texture, and combined feature spaces.

is given a separate score, as discussed in Sec. II D. In practice, radiologists read different views of the same patient together. To simulate this condition, we combined the discriminant scores of different views of the same mass from the same year to obtain a single case-based score for each mass. This analysis resulted in 127 average scores for 102 patients, because some patients had mammograms spanning multiple years or from both breasts, and masses in different breasts or from different years were averaged separately. As described in Sec. II D, we compared using either the maximum malignancy score or the average malignancy score as the combination method. These two methods both resulted in ROC curves with $A_z=0.91$. The distribution of the case-based scores using the averaging method is shown in Fig. 12. The ROC curves for film-based classification ($A_z=0.87 \pm 0.02$) and case-based classification ($A_z=0.91 \pm 0.02$) are shown in Fig. 13.

IV. DISCUSSION

Our results indicate that accurate segmentation of mammographic masses and the use of morphological features can

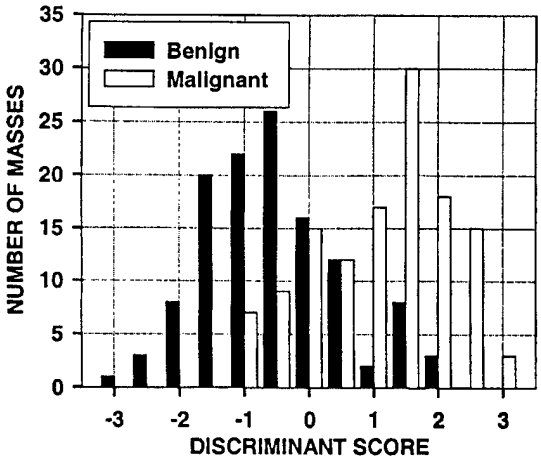


FIG. 11. The distribution of the film-based discriminant scores for leave-one-case-out classification of malignant and benign masses, using the combined feature space. The score of a mass on each film is considered independently.

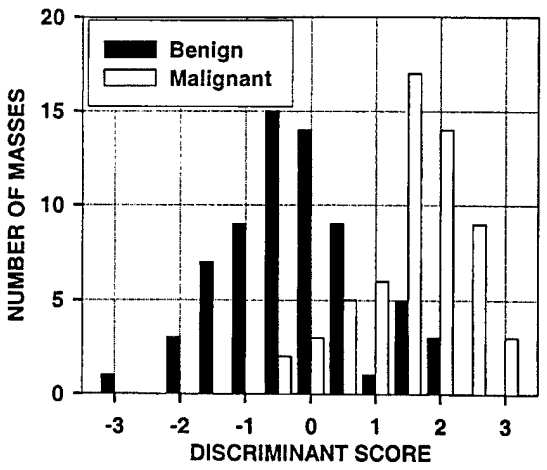


FIG. 12. The distribution of the case-based discriminant scores for leave-one-case-out classification of malignant and benign masses, using the combined feature space. The scores from the same mass of the same year have been averaged into a single score for the mass.

be effective in classifying breast masses as malignant or benign. When tenfold cross validation was used for data partitioning, the average classification accuracy with morphological features alone was equal to that with texture features alone ($A_z=0.85$). The average classification accuracy improved to $A_z=0.89$ when texture and morphological features were combined. In the tenfold cross-validation method, the test A_z values for each partition were computed separately. This meant that there were, on average, 24.9 films in each test partitioning. Due to the small number of cases used for computing the test ROC curves, the standard deviations of the A_z values were large, relative to those obtained using the leave-one-out method, as observed from Table III. As a result, the difference between the classifiers trained with the three different feature spaces did not reach statistical significance for any of the ten partitions shown in Table III. For the leave-one-case-out method, the scores from all ROIs were accumulated for the ROC analysis, as explained previously. This meant that the classification scores for all films were analyzed to obtain the test ROC curve. In this case, the classifier based on the combined feature space was significantly

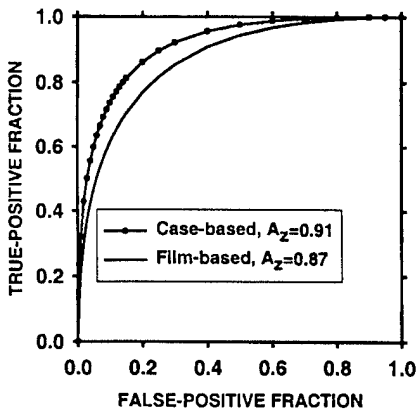


FIG. 13. Case-based and film-based ROC curves for classification of malignant and benign masses.

more accurate than that based on the texture feature space alone ($p=0.04$). The difference between the classifiers based on the combined and morphological feature spaces did not reach statistical significance.

We previously introduced a rubber-band straightening transform to analyze the margin characteristics of a mass in a texture feature space.[10] In this work, we developed a new three-stage segmentation method that consists of clustering, active contours, and spiculation detection; and evaluated the effectiveness of combining the morphological features extracted from the segmented mass and texture features for improving computerized breast mass classification. The morphological features used in this study were not novel;[5,26,28] and we had previously attempted to combine these features with texture features. However, with our previous mass segmentation method, we were unable to improve our texture-based classification results by including morphological features. This is a strong indication that the quality of segmentation is very important for morphological feature extraction.

The three-stage segmentation method used in this study adds two new stages to our previous segmentation method.[10] Previously, the clustering method was successful in segmenting the main portion of the mass from the background. However, one major limitation of clustering-based segmentation is that, even for well-circumscribed masses, the segmented shape contains many irregularities due to structured or random noises [see Fig. 4(d)]. Another limitation is that, to prevent merging with neighboring structures, the clustering parameters have to be chosen so that the segmented object is slightly smaller than the object that would visually be determined for a majority of the masses. Morphological features extracted from such a segmented mass may not adequately characterize the true morphology of the mass. The first new segmentation component of this study is the use of an active contour model for refining the clustering-based segmentation results. The second new component is the use of image gradient directions for detecting and segmenting spiculations. As shown in Fig. 9, the spiculation detection variable designed in this study was able to provide some separation between the spiculated and the nonspiculated masses. When the spiculation detection variable was used as the decision variable to classify the masses as spiculated or nonspiculated, the area A_z under the ROC curve was 0.85. However, this variable could not be directly used for the classification of the masses as malignant or benign, because almost half (64/127) of the malignant masses were visually characterized as nonspiculated by a radiologist experienced in mammographic interpretation.

The ability of each morphological feature to discriminate between the ROIs containing malignant and benign masses is shown in Table I in terms of the area A_z under the ROC curve. The A_z values indicate the accuracy of classifying the individual 249 ROIs as malignant or benign. The feature with the highest classification accuracy was the Fourier descriptor (FD). The stepwise method selected FD for all of the ten partitions shown in both the first and the last columns of Table III. When feature selection was performed using the morphological features alone, the contrast feature was selected, in addition to FD, for all of the ten partitions shown in the first column of Table III. The classification accuracy of the contrast feature is lower than those of several other features in Table I. However, contrast is the only feature in Table I that makes use of the gray scale information in the image. Therefore, compared to other morphological features, it seems to be able to introduce more complementary, and useful, information into the classifier when combined with FD.

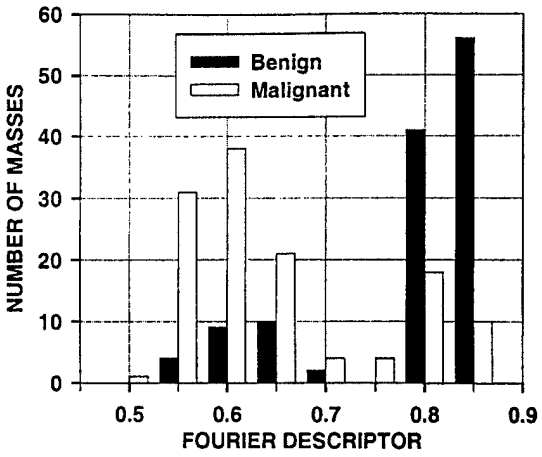


FIG. 14. The distribution of the Fourier descriptor feature for malignant and benign masses.

The ability of each texture measure to discriminate between malignant and benign masses is shown in Table II. It is observed that when used alone, the texture features are less effective than morphological features in classifying the masses in our data set. However, when texture features are combined using a linear classifier, the classification accuracy is comparable to classification using a linear classifier with morphological features alone. This may be an indication that the linear classifier is not as effective for combining these morphological features as for combining the texture features. We believe that a major reason for this is the distributions of the morphological features. It is known that the linear classifier is optimal for features with multivariate Gaussian class distributions with equal covariance matrices.[40] Due to the thresholding operation in segmentation (see the last paragraph of Sec. II B, and Fig. 3), the distributions of the morphological features in this study are very different from being Gaussian. As an example, the distribution of the Fourier descriptor feature is shown in Fig. 14. It can be observed that the distributions of both the benign and the malignant masses follow a bimodal distribution, very likely with the smaller peak corresponding to masses classified as spiculated, and the larger peak corresponding to those classified as nonspiculated. It is known that other types of classifiers, such as artificial neural networks or hybrid classifiers, perform better with non-Gaussian distributions. We will investigate the performance of other types of classifiers in these feature spaces in the future.

In a previous study, we had used the same texture features as those in this study, and had obtained an ROC area of 0.92

on a data set containing 238 masses.[4] The main reason for the lower accuracy with texture features in this study is the difference of the feature selection methods used in the two studies. In our previous study, the features were selected using the entire data set, as have been done in most studies in the CAD literature.[41–46] After feature selection, the data set was partitioned into training and test sets for formulation of the linear discriminant function. In the current study, both feature selection and classifier coefficient determination were performed on the training set. We have recently compared the effect of these two different approaches to feature selection on classifier performance prediction using a Monte Carlo simulation study.[39] We have found that, when feature selection is performed using the training set alone, the predicted test performance of the classifier is lower, in general, than that of a classifier trained with an infinite number of samples, as can be expected when a classifier is designed with a finite design sample set. However, when feature selection is performed using the entire set of available samples (training and test sets together), the predicted test performance can be higher or lower than that of a classifier trained with an infinite number of samples, depending on the number of available samples, the number of features, and the correlation between the features. The fact that the predicted performance of the classifier designed with a finite sample set can exceed that with an infinite sample set in the latter case indicates that feature selection using the entire available sample set can result in an overly optimistic prediction of the classifier performance. In studies with a clinical data set, there is no knowledge of the true class distributions, so it is difficult to predict which approach will be less biased. In order to provide a conservative prediction of the classifier performance for the general population, we chose to perform the feature selection on trainers alone in our current study.

Our data set contained 223 mammograms obtained less than six months before biopsy (preoperative mammograms) and 26 mammograms obtained more than six months prior to biopsy (prior mammograms). In order to obtain case-based average scores, we combined the scores from different years separately. Since the characteristics of the mass may change with time, combining scores across multiple years will not be reasonable. Similar to radiologists' interpretation,[4] case-based classification accuracy was higher than film-based accuracy, with $A_z=0.91$ and $A_z=0.87$ for the two methods, respectively.

An important feature of a CAD lesion classifier is its ability to alert radiologists to a suspicious lesion on a mammogram obtained at a time when the radiologist's suspicion level is not high enough to recommend biopsy. These prior mammograms, which are by definition more difficult to characterize, were included in our database because one would encounter such cases in clinical use or evaluation of a CAD system. If these 26 prior mammograms in our data set were excluded from the analysis, then case-based and film-based A_z values would be 0.94 and 0.88, respectively. Since the number of prior mammograms was small, we did not compare the classification accuracy of prior mammograms to that of preoperative mammograms in this study. When a larger set of prior mammograms is collected, it will be interesting and important to evaluate whether the computer classifier can predict the malignancy of the "unsuspected" masses in earlier years.

V. CONCLUSION

We have developed a fully automated three-stage segmentation method for delineation of mass boundary and detection and segmentation of spiculations. Morphological features describing the shape of the mass and texture features describing the margin characteristics of the mass were extracted from the segmented mass and a band of pixels surrounding the segmented mass, respectively. The data set was partitioned using a tenfold cross validation and a leave-one-case-out method for training and testing a classifier with stepwise feature selection followed by linear discriminant analysis. Using the combined feature space, the test classification accuracy was $A_z=0.89$ and $A_z=0.87$ for the tenfold cross validation and the leave-one-case-out methods, respectively. Case-based classification scores were obtained by averaging the test scores of the same mass from the same year. The area under the ROC curve for case-based classification was $A_z=0.91$. Our results indicate that combining morphological features extracted from the automatically segmented mass boundary with texture features can significantly improve the accuracy for computer-aided characterization of mammographic masses.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award (B.S.) from the USAMRMC No. (DAMD 17-96-1-6012), USPHS Grant No. CA 48129, and a Whitaker Foundation Grant (N.P.). The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC program.

[a]Electronic mail: berki@umich.edu

[1] G. Hermann, C. Janus, I. S. Schwartz, B. Krivisky, S. Bier, and J. G. Rabinowitz, "Nonpalpable breast lesions: Accuracy of prebiopsy mammographic diagnosis," Radiology **165**, 323–326 (1987).

[2] F. M. Hall, J. M. Storella, D. Z. Silverstond, and G. Wyshak, "Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography," Radiology **167**, 353 (1988).

[3] H. G. Jacobson and J. Edeiken, "Biopsy of occult breast lesions: Analysis of 1261 abnormalities," JAMA **263**, 2341–2343 (1990).

[4] H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: An ROC study," Radiology **212**, 817–827 (1999).

[5] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).

[6] S. Pohlman, K. A. Powell, N. A. Obuchowshi, W. A. Chilote, and S. Grundfest-Broniatowski, "Quantitative classification of breast tumors in digitized mammograms," Med. Phys. **23**, 1337–1345 (1996).

[7] R. M. Rangayyan, N. El-Faramawy, J. E. L. Desautels, and O. A. Alim, "Discrimination between benign and malignant breast tumors using a

region-based measure of edge profile acutance," in *Digital Mammography '96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996).

[8] J. L. Viton, M. Rasigni, G. Rasigni, and A. L. Llebaria, "Method for characterizing masses in digital mammograms," Opt. Eng. (Bellingham) **35**, 3453–3459 (1996).

[9] Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," Acad. Rad. **5**, 155–168 (1998).

[10] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1998).

[11] D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," Invest. Radiol. **23**, 240–252 (1988).

[12] J. Y. Lo, J. A. Baker, P. J. Kornguth, and C. E. Floyd, "Computer-aided diagnosis of breast cancer: Artificial neural network approach for optimized merging of mammographic features," Acad Radiol. **2**, 841–850 (1995).

[13] J. A. Baker, P. J. Kornguth, J. Y. Lo, and C. E. Floyd, "Artificial neural network: Improving the quality of breast biopsy recommendations," Radiology **198**, 131–135 (1996).

[14] J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon," Radiology **196**, 817–822 (1995).

[15] C. J. D'Orsi and D. B. Kopans, "Mammographic feature analysis," Seminars in Roentgenology **28**, 204–230 (1993).

[16] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," IEEE Trans. Med. Imaging **15**, 598–610 (1996).

[17] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).

[18] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: An artificial neural network with morphological features," Proc. World Cong. Neural Net. II , 876–879 (1995).

[19] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms," Med. Phys. **23**, 1671–1684 (1996).

[20] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models,," Int. J. Comput. Vis. **1**, 321–331 (1987).

[21] C. S. Poon and M. Braun, "Image segmentation by a deformable contour model incorporating region analysis," Phys. Med. Biol. **42**, 1833–1841 (1997).

[22] D. J. Williams and M. Shah, "A fast algorithm for active contours and curvature estimation," CVGIP: Image Understand. **55**, 14–26 (1992).

[23] H.-P. Chan, N. Petrick, and B. Sahiner, "Computer-aided breast cancer diagnosis" in *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*, edited by A. Jain, A. Jain, S. Jain, and L. Jain (World Scientific, 2000), Chap. 6.

[24] N. Karssemeijer and G. te Brake, "Detection of stellate distortions in mammograms," IEEE Trans. Med. Imaging **15**, 611–619 (1996).

[25] J. J. Koenderink and A. J. van Doorn, "Generic neighborhood operators," IEEE Trans. Pattern Anal. Mach. Intell. **14**, 597–605 (1992).

[26] N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," Med. Phys. **26**, 1642–1654 (1999).

[27] S. Mori, H. Nishida, and H. Yamada, *Optical Character Recognition* (Wiley, New York, 1999).

[28] L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications," IEEE Trans. Med. Imaging **13**, 263–274 (1994).

[29] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, G. M. M. , and D. D. Adler, "Classification of masses on mammograms using a rubber-band straightening transform and feature analysis," Proc. SPIE **2710**, 44–50 (1996).

[30] B. Sahiner, H. P. Chan, N. Petrick, G. M. M. , and M. A. Helvie, "Characterization of masses on mammograms: Significance of the use of the rubber-band straightening transform," Proc. SPIE **3034**, 491–500 (1997).

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).

[32] M. M. Galloway, "Texture classification using gray level run lengths," Comput. Graphics **4**, 172–179 (1975).

[33] C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat. Med. **17**, 1033–1053 (1998).

[34] C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance for differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging*, edited by F. Deconinck (Martinus Nijhoff, the Hague, 1984).

[35] N. R. Draper, *Applied Regression Analysis* (Wiley, New York, 1998).

[36] M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics* (SPSS, Chicago, IL, 1993).

[37] H. P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature space," Med. Phys. **25**, 2007–2019 (1998).

[38] N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," Med. Phys. **23**, 1685–1696 (1996).

[39] B. Sahiner, H. P. Chan, N. Petrick, R. F. Wagner, and L. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size," Med. Phys. **27**, 1509–1522 (2000).

[40] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990).

[41] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," Radiology **187**, 81–87 (1993).

[42] K. G. A. Gilhuijs and M. L. Giger, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," Med. Phys. **25**, 1647–1654 (1998).

[43] B. S. Garra, B. H. Krasner, S. C. Horri, S. Ascher, S. K. Mun, and R. K. Zeman, "Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis," Ultrason. Imaging **15**, 267–285 (1993).

[44] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," IEEE Trans. Med. Imaging **14**, 537–547 (1995).

[45] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Med. Phys. **19**, 1475–1481 (1992).

[46] Z. Huo, M. L. Giger, D. E. Wolverton, and W. Zhong, "Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection," Med. Phys. **27**, 4–12 (2000).

Computer-Aided Characterization of Mammographic Masses: Accuracy of Mass Segmentation and its Effects on Characterization

Berkman Sahiner
Nicholas Petrick
Heang-Ping Chan
Lubomir M. Hadjiiski
Chintana Paramagul
Mark A. Helvie
Metin N. Gurcan

Department of Radiology
University of Michigan, Ann Arbor

Correspondence:
Berkman Sahiner, Ph.D.
Department of Radiology
University of Michigan
1500 E. Medical Center Drive
CGC B2102
Ann Arbor, MI 48109-0904
Telephone: (734) 647-7429
Fax: (734) 647-8557
e-mail: berki@umich.edu

Abstract

Mass segmentation is used as the first step in many computer-aided diagnosis (CAD) systems for classification of breast masses as malignant or benign. The goal of this work was to study the accuracy of an automated mass segmentation method developed in our laboratory, and to investigate the effect of the segmentation stage on the overall classification accuracy. The automated segmentation method was quantitatively compared with manual segmentation by two expert radiologists (R1 and R2) using three similarity or distance measures on a data set of 100 masses. The area overlap measures between R1 and R2, the computer and R1, and the computer and R2 were 0.76±0.13, 0.74±0.11, and 0.74±0.13, respectively. The inter-observer difference in these measures between the two radiologists was compared to the corresponding differences between the computer and the radiologists. Using three similarity measures and data from two radiologists, a total of six statistical tests were performed. The difference between the computer and the radiologist segmentation was significantly larger than the inter-observer variability in only one test. Two sets of texture, morphological, and spiculation features, one based on the computer segmentation, and the other based on radiologist segmentation, were extracted from a data set of 249 films from 102 patients. A classifier based on stepwise feature selection and linear discriminant analysis was trained and tested using the two feature sets. The leave-one-case-out method was used for data sampling. For case-based classification, the area A_z under the receiver operating characteristic (ROC) curve was 0.89 and 0.88 for the feature sets based on the radiologist segmentation and computer segmentation, respectively. The difference between the two ROC curves was not statistically significant.

Keywords: computer-aided diagnosis, mammography, breast masses, classification, segmentation

I. INTRODUCTION

Mammography is currently the most sensitive method to detect early breast cancer. However, many suspicious findings on mammograms are benign. The most important mammographic signs of malignancy are masses and microcalcifications. The benign biopsy rate for mammographically suspicious, nonpalpable breast masses is about 70 to 80% [1-3]. In order to reduce patient anxiety and morbidity, as well as to reduce costs, it is important to reduce the biopsy rate without missing malignancies. Besides its potential benefits in detecting mammographic lesions that might otherwise be overlooked, computer-aided diagnosis (CAD) also has the potential to increase the efficacy of mammography by assisting the radiologists to better differentiate malignant and benign lesions and reduce the benign biopsy rate.

Many CAD systems use a three-stage approach for classification of breast masses as malignant or benign. First, the mass is segmented from the background tissue. Next, features relevant to the classification task are extracted from the mass and the surrounding tissue. Finally, the feature values are merged into a malignancy score using a classifier. The segmentation methods used for classification of mammographic masses can be manual [4, 5], semi-automated [6, 7], or fully automated [8, 9]. Mudigonda *et al.* [4] used hand-segmented mass boundaries to extract gradient-based features and texture measures from a ribbon surrounding the mass, and used the features in a stepwise discriminant analysis classifier. Bruce *et al.* [5] used multiresolution features to quantify mass shapes that were segmented manually by a radiologist. Pohlman *et al.* [6] segmented masses using an adaptive region growing algorithm. If a mass could not be segmented after repeated manual adjustments of the region growing algorithm, it was excluded from the data set. They found that in the task of differentiating invasive cancer and benign lesions, their tumor boundary roughness feature achieved a

classification accuracy comparable to those of two experienced radiologists who specialized in mammography. Kilday *et al.* [7] extracted mass shapes using interactive gray-level thresholding. Masses that were not successfully segmented were excluded from analysis. Morphological features and patient age were used to classify the masses into cancer, cyst, and fibroadenoma categories. Huo *et al.* [8] used a fully automated region growing algorithm with a multiple transition point technique to segment the masses. Features extracted from the margin and the density of the masses were used in a two-stage hybrid classifier consisting of a rule-based stage and an artificial neural network stage. The hybrid classifier achieved an area A_z under the receiver operating characteristic (ROC) curve of 0.94 for a data set of 65 cases. On an independent data set of 110 cases, their classifier had an A_z of 0.82 [10]. Sahiner *et al.* used a clustering algorithm for fully automated segmentation. Texture features were extracted from a band of pixels surrounding the mass which had been transformed to a rectangular strip using the rubber-band straightening transformation (RBST) [9]. A linear discriminant analysis (LDA) classifier was used for the classification task. In an ROC study that investigated the effect of this classifier on radiologists' classification of breast masses, Chan *et al.* [11] showed that the computer classifier could significantly improve radiologists' accuracy in characterization of masses.

Hand-segmentation by a radiologist may assure that most of the mass boundaries are correctly delineated. However, hand-segmentation can be time-consuming and subject to intra- and inter-observer variations. A fully automated segmentation algorithm has the advantages of speed and reproducibility. However, it is very difficult to develop an automated method that can accurately and consistently detect mass boundaries because many masses have ill-defined boundaries and often overlap with fibroglandular tissue.

This study has two major goals. The first goal is to quantitatively compare an automated segmentation method developed in our laboratory with manual segmentation by two expert radiologists (R1 and R2), and to analyze the measure of agreement between R1, R2, and the computer. Since there is no ground truth of the mass boundaries and radiologists exhibit variations in mass boundary delineation, it will be more meaningful to determine if the difference between mass boundaries delineated by the computer and a radiologist falls within the range of variation between radiologists, rather than to measure the absolute deviation of the computer-segmented boundaries from any one of the radiologists. The second goal is to compare the classification accuracy based on computer segmentation to that based on manual segmentation by a radiologist. This comparison is intended to give us an indication as to whether future characterization work needs to focus on better mass segmentation or on improved features and better classifier design. To our knowledge, neither of these goals has previously been pursued in the literature.

In the next Section, we first describe the data set used in this study. Our fully automated mass segmentation method based on an active contour (AC) model is described in Section II.B. The similarity measures used for quantifying the agreement between the segmentations by R1, R2, and the computer are discussed in Section II.C, while Sections II.D and II.E briefly describe the feature extraction and classification techniques used for characterizing the masses as malignant or benign. The quantitative comparisons and the classification accuracy results are presented in Section III, with a discussion of the results and conclusions in the following two sections.

II. METHODS

A. Data Set

The mammograms used in this study were randomly selected from the files of patients in the Radiology Department at the University of Michigan who had undergone biopsy. All mammograms were acquired with dedicated mammographic systems. The criteria for inclusion of a mammogram in the data set were that the mammogram contained a biopsy-proven mass, and that approximately equal numbers of malignant and benign masses were present in the data set.

Our data set consisted of 249 mammograms from 102 patients. The mammograms contained a total of 122 benign and 127 malignant masses. The histology of the masses were determined by biopsy and pathologic analysis. The probability of malignancy of the biopsied mass on each mammogram was ranked by a Mammography Quality Standards Act (MQSA) approved radiologist on a scale of 1 (most benign mammographic appearance) to 10 (most malignant mammographic appearance). The distribution of the malignancy ranking of the masses on each view is shown in Fig. 1. The malignant and benign masses overlap over the entire range of suspicion for malignancy, indicating that it was difficult for the radiologists to distinguish malignant from benign masses in this data set. This is consistent with the fact that all masses underwent biopsy after clinical evaluation of the cases. The size of the masses in our data set ranged from 5 to 29 mm (mean size=12.5 mm). The distributions of the sizes for the malignant and benign masses are shown in Fig. 2. It is observed that the lesion size distribution for malignant masses is similar to that for benign masses.

The mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel size of 100 μm x 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units,

with a slope of 0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually, with the OD range extending to 3.5.

The location of the biopsied mass was identified by a radiologist, and a region of interest (ROI) containing the mass was extracted. The size of the ROI was chosen such that the radiologist-marked lesion and a band of about 70-pixel-wide surrounding background were included in the ROI. R1 was requested to manually delineate all masses in the data set, and R2 was requested to delineate a randomly selected subset of 100 masses. Both radiologists used specially designed interactive software on a display workstation to outline the masses. After initial experimentation, it was decided that, for spiculated masses, only the central tumor would be delineated by the radiologists, not the spiculations. The main reason for this approach is that we observed that the decision to include or exclude a spiculation by the radiologists was very subjective. The visibility of individual spiculations for a mass can range from obvious to very subtle and ill-defined. When the number of subtle spiculations was large, it was practically impossible for the radiologists to individually trace them. When these spiculations were not traced, it was not known if the radiologist did not consider them to be spiculations or if they preferred not to trace them individually. Our mass classification techniques depend largely on the segmentation of the tumor. Morphological features are extracted from the central tumor and texture and spiculation features extracted from the tissue around the mass margin. The additional variability due to the uncertainty in outlining the spiculations by radiologists may therefore mask the goodness of the segmentation for our purpose.

Segmentation of masses that overlap with dense breast tissue is in general more difficult, because mass boundaries may become indistinct or occluded in such cases. The inter-observer agreement on mass boundaries therefore depends on the breast density. The breast density

histogram for the mammograms in our data set, as estimated by R2 using the American College of Radiology [12] recommended BI-RADS lexicon is shown in Fig. 3.

B. Mass Segmentation

Initial mass segmentation

The mass segmentation method employed in this study starts with the initial detection of a mass shape within an ROI using a pixel-by-pixel K-means clustering algorithm, which has been discussed in detail in the literature [13, 14]. Clustering segments one or more disjoint objects within the ROI. If more than one object is segmented, the largest connected object is selected. The selected object is then filled, grown in a local neighborhood, and eroded and dilated with morphological operators. The implementation details of these steps have been described elsewhere [9].

Active contour segmentation

Although initial mass segmentation resulted in reasonable mass shapes for most of the masses, further refinement was necessary before detection and segmentation of the spiculations. We used an active contour (AC) model for mass shape refinement.

An AC is a deformable continuous curve, whose shape is controlled by internal forces (the model, or *a-priori* knowledge about the object to be segmented) and external forces (the image) [15]. The internal forces impose a smoothness constraint on the contour, and the external forces push the contour towards salient image features, such as edges. To solve a segmentation problem, an initial boundary is iteratively deformed so that the energy due to internal and external forces is minimized along the contour. An AC model was used for mass segmentation in the mass detection algorithm in [16], but the quality of the segmentation was not assessed.

The internal energy components used in our AC model were the continuity and curvature of the contour, as well as the homogeneity of the segmented object. The external energy components were the negative of the smoothed image gradient magnitude, and a balloon force that exerted pressure at a normal direction to the contour. The contour was represented by the vertices of an N-point polygon whose vertices were $v(c)=(x_c,y_c)$, $c=1,...,N$. The energy to be minimized was defined as

$$E = \sum_{c=1}^{N} \left[w_{curv} E_{curv}(c) + w_{cont} E_{cont}(c) + w_{grad} E_{grad}(c) + w_{bal} E_{bal}(c) \right] + w_{hom} E_{hom} \qquad (1)$$

where *curv, cont, grad, bal,* and *hom* stand for curvature, continuity, gradient, balloon force, and homogeneity, respectively, and each energy term is associated with a weight, w.

The curvature energy term is represented by an approximation to the second derivative of the contour, $E_{curv}(c) = \left\| \mathbf{v}(c-1) - 2\mathbf{v}(c) + \mathbf{v}(c+1) \right\|$. This term is large when the angle between the two sides of the polygon that meet at vertex $v(c)$ is small (Fig. 4). By discouraging small angles at vertices, this term attempts to smooth the contour. The continuity energy term, $E_{cont}(c)$, is represented by the deviation of the length of the line segment s_c between vertices $v(c)$ and $v(c+1)$ from the average line segment length $\bar{s} = \sum s_c / N$. Therefore this term tries to maintain regular spacing between the vertices along the contour. The image gradient magnitude is obtained by smoothing the image with a low-pass filter, finding the partial derivatives in the horizontal and vertical directions, and then computing the magnitude of the partial derivative vector. Since the gradient energy, $E_{grad}(c)$, is defined as the negative of the gradient magnitude, minimizing this term attracts the contour to the object edges. The effect of the balloon energy, $E_{bal}(c)$, can best be understood by considering how this energy component is affected when the vertex $v(c)$ is moved to a new location $v'(c)$ during the energy minimization process described

below. The normal direction to the contour at vertex $v(c)$ is defined as the average of the normals to the two sides of the polygon that meet at vertex $v(c)$. The balloon energy is defined as the cosine of the angle between this normal vector and the vector $v'(c)$-$v(c)$. If the weight term w_{bal} is negative, then this energy component encourages the contour to expand in the normal direction. The balloon energy term is required to prevent the contour from collapsing onto itself, which is a well-known phenomenon in AC models [17]. The purpose of the homogeneity energy term, $E_{hom}(c)$, is to make the object and the background regions defined by the area inside and outside of the contour as homogeneous as possible, and to maximize the difference between the two regions. In our study, the background region is defined as the union of pixels that are within a distance d from the object region, where d is chosen so that the area of the background region is the same as that of the object. The homogeneity energy is defined as the ratio of within-region sum-of-squares and between-region sum-of-squares, where the two regions are the object and the background regions. The precise definition of these sum-of-squares terms can be found in the literature [18]. In an image where both the object and the background have uniform but different gray-level values, the homogeneity energy will be zero when the contour is optimized.

To minimize the contour energy, we used a greedy algorithm that was first proposed by Williams and Shah [19]. In this algorithm, the contour is iteratively optimized, starting with the initial contour provided by the output of the first stage segmentation. At each iteration, a neighborhood of each vertex is examined, and the vertex is moved to the location that minimizes the contour energy. The algorithm stops when there is no movement of the vertices, or when all the vertices of the contour are at locations already visited at a previous iteration.

C. Distance and Similarity Measures

We considered two distance measures and one similarity measure to quantify the difference between two segmented objects. The distance measures were defined in terms of the minimum Euclidean distance (MED) between a point p and a curve B in the Cartesian plane. If the curve B is described in terms of q points $\{b_1,\ldots,b_q\}$, then $MED(p,B)$ is defined as

$$\mathrm{MED}(p,B) = \min_{i\in\{1,..,q\}} \|p - b_i\|. \tag{2}$$

Let $A=\{a_1,\ldots,a_p\}$ and $B=\{b_1,\ldots,b_q\}$ denote the two closed contours to be compared. The directed Hausdorff distance $h(A,B)$ identifies the point a in A that is farthest from the curve B, and measures the distance between a and B

$$h(A,B) = \max_{i\in\{1,\ldots,p\}} \{\mathrm{MED}(a_i,B)\} \tag{3}$$

The Hausdorff distance [20] (illustrated graphically in Fig. 5) is the first distance measure used in this paper, and it is defined in terms of the directed Hausdorff distance as

$$H(A,B) = \max\{h(A,B), h(B,A)\} \tag{4}$$

It is well known that the Hausdorff distance is a metric, i.e., it satisfies the identity and symmetry equalities and the triangle inequality, and is non-negative [20]. It does not require an explicit pairing of points between A and B. One disadvantage of the Hausdorff distance is that it does not measure how much A and B are dissimilar *on the average*. For example, even when the two closed contours are identical at all points except one, the Hausdorff distance can be large. We therefore defined a second distance measure, the average minimum Euclidean distance (AMED), by averaging the average distance of a_i to B and the average distance of b_i to A:

$$\mathrm{AMED}(A,B)=\frac{\sum_{i=1}^{p} MED(a_i,B)}{2p}+\frac{\sum_{i=1}^{q} MED(b_i,A)}{2q} \tag{5}$$

The similarity measure we employed was the area overlap measure (AOM), as illustrated in Fig. 5. The AOM between two closed contours A and B is defined as

$$\mathrm{AOM}(A,B)=\frac{Area\{S_A \cap S_B\}}{Area\{S_A \cup S_B\}}, \tag{6}$$

where S_A and S_B are the interior regions of A and B, respectively. It is easily seen that AOM(A,B)=1 when A and B are identical, AOM(A,B)=0 when S_A and S_B do not intersect, and $1 \geq \mathrm{AOM}(A,B) \geq 0$ when the similarity of the two closed contours are between these two extremes.

D. Feature Extraction

Morphological features

Thirteen morphological features are extracted from the segmented masses. The first five morphological features were based on the normalized radial length (NRL), defined as the Euclidean distance from the object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object [7]. In our previous studies, we found that NRL mean, standard deviation, entropy, area ratio, and zero crossing count were useful for discriminating between objects containing masses and normal tissue [21]. The next six features were the perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast of the object. The definition of these features can be found in the literature [21]. These features were also shown to be useful in discriminating between objects containing masses and normal tissue [21].

The twelfth feature, convexity, was defined as the ratio of the area of the segmented object to the area of the smallest convex object that contained the object. If the object was convex, as was

the case with many benign masses, then this feature would approach its maximum value of unity. If the object shape was highly non-convex, as was the case with many malignant masses, then the value of this feature would be small.

The last feature was the summary Fourier descriptor measure [22], which was based on the Fourier transform of the object boundary sequence. Objects with irregular contours have more high-frequency components than those with smooth contours [23]. The Fourier descriptor measure therefore contains potentially useful information for discriminating between benign and malignant masses.

Texture features

The texture of the region surrounding the mass can yield important features for its classification. Since spiculations and the gradient of the opacity caused by the mass are approximately radially oriented, the texture of the region surrounding a mass is expected to have a radial dependence. However, most texture extraction methods are designed for texture orientations in a uniform direction (horizontal, vertical, or at a certain angle between these two directions). To be able to extract meaningful texture features from the region surrounding a mass, we have designed a rubber band straightening transform (RBST) that maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region) [9, 24, 25]. The width of the band was chosen as 4 mm. In the transformed image, the border of the mass is expected to appear approximately as a horizontal edge, and spiculations are expected to appear approximately as vertical lines.

The texture features extracted from the RBST images include 13 texture measures, each calculated at 4 directions and 10 distances, from the spatial gray-level dependence (SGLD) matrices and 20 run-length statistics (RLS) features, as described in our previous work [9]. The

definition of these features [26, 27] and the parameters used in this study can be found in the literature [9].

Spiculation features

Three spiculation features are extracted based on whether points (i_c,j_c) on the mass contour lie on the path of a spiculation. Since a spiculation is a linear structure, the image gradients at different points that lie on the same spiculation have similar phase directions. We have previously defined a spiculation measure in terms of the statistics of these phase directions [28]. This spiculation measure, described in detail in the literature [28], is briefly discussed next, and the three spiculation features defined in terms of the spiculation measure are introduced.

Let c denote a pixel on the mass boundary determined by the segmentation algorithm, $1 \le c \le N$, where N is the total number of points on the boundary, and let (i_c,j_c) be the Cartesian coordinates of the mass boundary pixel c. If (i_c,j_c) lies on the path of a spiculation, and the image pixel (i,j) is part of this spiculation, then the image gradient at (i,j) will be more or less perpendicular to the line joining (i_c,j_c) to (i,j). We make use of this property by defining the spiculation measure $x(i_c,j_c)$ at the mass boundary pixel c as

$$x(i_c, j_c) = \frac{1}{N_s}\left(\sum_{(i,j)\varepsilon S(i_c, j_c)} \theta(i,j) \right), \tag{7}$$

where $0 \le \theta \le \pi/2$ is the obtuse angle between the gradient phase at (i,j) and the line joining (i_c,j_c) and (i,j), $S(i_c,j_c)$ is an ROI, and N_s is the number of pixels in this ROI. The spiculation measure is therefore defined as the average of θ in $S(i_c,j_c)$. The ROI $S(i_c,j_c)$ is defined based on a-priori knowledge about spiculations, such as the fact that they lie outside the mass, have a positive contrast, and are generally within a radial sector of $\pm\pi/4$ centered about the normal to

the mass contour at (i_c,j_c) [28, 29]. The computation of the image gradient is based on convolution with Gaussian derivatives [30].

The spiculation measure is computed for all points $c=1,...,N$ on the mass boundary. Three features are derived from the spiculation measure to quantify the degree of spiculation of a mass. The first feature is the average of the spiculation measure for all pixels on the mass boundary. The second feature is the percentage of boundary pixels with a spiculation measure larger than $\pi/4$, and the third feature is the average of the spiculation measure for those pixels with a spiculation measure larger than $\pi/4$.

E. Classification

The data was partitioned into trainers and testers using a leave-one-case-out methodology. In the leave-one-case-out method, all patients except one serve as trainers in the classifier design stage for a given partition. The designed classifier is then applied to the case that is left out to obtain test discriminant scores for the images of that patient. By leaving each patient out in a round-robin order, test discriminant scores are obtained for images of all patients. The classifier design stage consisted of stepwise feature selection and formulation of the linear discriminant function using the selected features as the predictor variables. Since there were 102 patients in our data set, the stepwise feature selection and LDA design processes were performed a total of 102 times, using a different training set defined by the leave-one-case-out method each time. The principles of LDA with stepwise feature selection [31, 32] and their application to CAD [9, 33-35] can be found in the literature.

The discriminant scores were analyzed with ROC methodology, using the LABROC program of Metz *et al.* [36]. The classification accuracy was measured by the area A_z under the ROC

curve. The statistical significance of the difference between the ROC curves obtained under different conditions was tested using the CLABROC program of Metz *et al.*[37].

III. RESULTS

A. Segmentation

The mass boundaries obtained by the AC model, and radiologists R1 and R2 were compared to one another in terms of the Hausdorff distance, the AMED, and the AOM for the 100 masses that were segmented by both radiologists. Table I shows the mean and standard deviation of these measures computed using the three possible paired comparisons of the segmentations (R1-R2, AC-R1, and AC-R2). Figures 6-8 show the distribution of the Hausdorff distance, the AMED, and the AOM, respectively, for the three comparisons. Figure 9 shows the segmented contours by AC, R1, and R2 for three masses. In terms of the area overlap measure, the example in the left column shows a good segmentation (average AOM = AOM(AC,R1)/2 + AOM(AC,R2)/2 = 0.92), the example in the middle column shows an average segmentation (average AOM=0.73), and that in the right column shows a poor segmentation (average AOM=0.50).

In order to test whether the agreement between the two radiologists was statistically different from that between the active contour and R2, we performed a paired t-test between the measures computed using the R1-R2 pairing and the AC-R2 pairing (Table II). The Hausdorff distance and the AMED computed using the R1-R2 pairing were not statistically different from that using the AC-R2 pairing ($p=0.07$ and 0.09, respectively). The AOM was significantly different ($p=0.02$). The paired t-tests comparing the measures computed using the R1-R2 pairing and the AC-R1 pairing showed no significant difference for any of the three measures ($p=0.52$, 0.34, and 0.10 for the Hausdorff distance, the AMED and the AOM, respectively). Thus, out of six

statistical tests, only one comparison indicated that the difference between the computer and the radiologist segmentation was significantly higher than that between the two radiologists.

B. Classification

All features used in this study depend either explicitly or implicitly upon mass segmentation. We were therefore interested in comparing classification results using features extracted from the masses based on computer segmentation to those based on radiologist's segmentation. Using the data partitioning and classification technique described in Section II.E, the A_z value for the 249 masses manually segmented by R1 was 0.86. The A_z value for the automated active contour segmentation method was also 0.86. These two ROC curves are plotted in Fig. 10. We also averaged the discriminant scores of different views of the same mass from the same examination to obtain a single case-based score for each mass (per-case analysis). This analysis resulted in 127 average scores for 102 patients because some patients had mammograms spanning multiple years or from both breasts, and masses in different breasts or from different examinations were averaged separately. The A_z values for the computer segmentation and the segmentation by R1 were 0.88 and 0.89, respectively. These case-based ROC curves are also plotted in Fig. 10. The difference in the ROC curves between the computer and the radiologist segmentation was not statistically significant ($p>0.5$).

IV. DISCUSSION

The quantitative comparison of the distance and similarity measures in Table I reveals that, on average, the agreement on the segmented mass boundaries between the two radiologists was consistently higher than that between the computer and R1 or R2. However, the numerical differences in these measures for the three paired comparisons were not large. For example, in

terms of AOM, the agreement between the two radiologists was 0.76±0.13, whereas that between the computer and R1 was 0.74±0.11, and that between the computer and R2 was 0.74±0.13.

The histograms shown in Fig. 8 indicate that for any of the three possible paired comparisons (R1-R2, AC-R1, or AC-R2) more than 85% of the AOM values were between 0.6 and 0.9. It is also interesting to note that the AOM(R1,R2) rarely (only 7 out of 100 masses) exceeded 0.90. The average minimum Euclidean distance was less than 1.0 mm for all three paired comparisons, and the Hausdorff distance was less than 3.0 mm. Since the RBST was applied to a 4-mm-wide band around the segmented object, this result indicates that, for an average mass, most of the band contained the intended image region.

Despite the fact that the average difference between AOM(R1,R2) and AOM(AC,R2) was small (ΔAOM=0.02) relative to the standard deviations, the difference was statistically significant. The explanation for this finding lies in the high correlation between AOM(R1,R2) and AOM(AC,R2) (Pearson's r=0.73). Essentially, this meant that even for cases for which AOM(R1,R2) was low, AOM(AC,R2) was slightly, but more or less consistently, lower. The other five statistical tests did not achieve statistical significance, although the p values were small (less than 0.10) for two of them.

The Hausdorff distance has the nice property of satisfying the axioms of a metric. However, it is based on a maximum distance principle, and does not directly depend on how close the two segmented boundaries are on the average. In an effort to include measures that provide an indication of the average agreement (or disagreement) between the segmented boundaries, we also used the AMED and AOM measures, although they do not satisfy the requirements of a metric. Other measures, such as the root mean-squared radial distance between the boundaries [38] are also possible. Neither the Hausdorff distance nor the AMED are scale invariant. This

means that a large segmentation error (in terms of actual distance in millimeters) results in a large Hausdorff distance (or AMED) regardless of the size of the mass. In practice, an N-mm error may be more tolerable for a large mass than for a small mass. By normalizing the area overlap by the union of the areas, the AOM achieves scale invariance. Despite these differences, the three measures in this study were correlated. The average correlation (averaged over the three possible paired comparisons) between the Hausdorff distance and AMED was 0.91. AOM was negatively correlated with the two other distance measures because it is a measure of similarity. The average correlation between AOM and AMED was –0.76, and that between AOM and the Hausdorff distance was –0.64. The high correlation between the Hausdorff distance and AMED indicates that for the task of comparing the segmentation of the central mass on mammograms, the two could be used interchangeably.

The overall classification accuracy A_z using features extracted based on radiologist segmentation and computer segmentation were nearly identical. This may indicate that the segmentation algorithm for outlining the central tumor in our classification method is satisfactory and that the features used in our current classifier are robust with respect to the small inaccuracies in our automated mass segmentation algorithm.

In an effort to further investigate how the agreement between the computer and the radiologist affects the classification accuracy, we studied the classification accuracy with the 50% of the masses that had the lowest AOM(AC,R1) value (N=124, average AOM=0.62). Presumably, these were the masses for which the computer segmentation performed relatively poorly. We would expect to see a lower classification accuracy using features based on the computer segmentation with these masses compared to the accuracy with features based on the radiologist segmentation. The A_z values with the computer segmentation and radiologist segmentation for

these 124 masses were 0.83 and 0.82, respectively. We therefore observed no decrease in classification accuracy with the computer segmentation, even when we analyze only the masses for which the radiologist and the computer had a high disagreement. The classification accuracy for these masses was lower than those that fell into the upper 50% in terms of AOM (A_z=0.88 and 0.89 using computer and radiologist segmentation, respectively). This may indicate that the masses with low AOM values may be generally more difficult to classify than those with high AOM. The radiologist's malignancy ratings for the masses (Fig. 1) seem to be consistent with this observation. The average malignancy ratings of the malignant and benign masses for the lower 50% AOM were 5.83 and 2.37 respectively, compared to 6.81 and 2.00 for the upper 50%. The lower malignancy ratings for malignant masses and the higher ratings for benign masses in the lower 50% group suggest that these masses are more difficult to classify.

V. CONCLUSION

We have quantitatively analyzed the accuracy of our automated mammographic mass segmentation algorithm by comparing its performance with manual segmentation by two radiologists. The average minimum Euclidean distance between the segmented boundaries by the computer and the radiologists was less than 1.0 mm, and the Hausdorff distance was less than 3.0 mm, on the average. The inter-observer difference between the two radiologists was compared to the differences between the computer and the individual radiologist segmentations using six paired t-tests on different distance and similarity measures. The difference between the computer and the radiologist segmentation was significantly higher than that between the two radiologists in only one out of the six comparisons. Two classifiers were trained and tested for classifying the masses as malignant or benign, one based on features extracted from the computer-segmented masses, and the other based on features extracted from the radiologist-segmented masses. The

two classifiers were nearly identical in terms of the area A_z under the ROC curve. These results suggest that the segmentation method for outlining the central tumor in our mass classification algorithm may be satisfactory. Our future research efforts will focus on designing new features and improving the classification methods in our mass characterization algorithm.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award from the USAMRMC (DAMD 17-96-1-6012) (B.S.), a USPHS Grant CA 48129, a USAMRMC grant (DAMD 17-96-1-6254), and a Career Development Award from the USAMRMC (DAMD 17-98-1-8211) (L.H.) . The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC program.

REFERENCES

[1] G. Hermann, C. Janus, I. S. Schwartz, B. Krivisky, S. Bier, and J. G. Rabinowitz, "Nonpalpable breast lesions: Accuracy of prebiopsy mammographic diagnosis," *Radiol.*, vol. 165, pp. 323-326, 1987.

[2] F. M. Hall, J. M. Storella, D. Z. Silverstond, and G. Wyshak, "Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography," *Radiol.*, vol. 167, pp. 353, 1988.

[3] H. G. Jacobson and J. Edeiken, "Biopsy of occult breast lesions: Analysis of 1261 abnormalities," *JAMA*, vol. 263, pp. 2341-2343, 1990.

[4] N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels, "Gradient and texture analysis for the classification of mammographic masses," *IEEE Trans. Med. Img.*, vol. 19, pp. 1032-1043, 2000.

[5] L. M. Bruce and R. R. Adhami, "Classifying mammographic mass shapes using the wavelet transform modulus-maxima method," *IEEE Trans. Med. Img.*, vol. 18, pp. 1170-1177, 1999.

[6] S. Pohlman, K. A. Powell, N. A. Obuchowshi, W. A. Chilote, and S. Grundfest-Broniatowski, "Quantitative classification of breast tumors in digitized mammograms," *Med. Phys.*, vol. 23, pp. 1337-1345, 1996.

[7] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," *IEEE Trans. Med. Img.*, vol. 12, pp. 664-669, 1993.

[8] Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Rad.*, vol. 5, pp. 155-168, 1998.

[9] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Med. Phys.*, vol. 25, pp. 516-526, 1998.

[10] Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness," *Acad. Rad.*, vol. 7, pp. 1077-1084, 2000.

[11] H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. S. Gopal, "Improvement of radiologists'

characterization of mammographic masses by computer-aided diagnosis: an ROC study," *Radiol.*, vol. 212, pp. 817-827, 1999.

[12] *American College of Radiology. Breast Imaging - Reporting and Data System (BI-RADS)*, Third Edition ed. Reston, VA: American College of Radiology, 1998.

[13] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: An artificial neural network with morphological features," *Proc. World Cong. Neural Net.*, vol. II, pp. 876-879, 1995.

[14] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms," *Med. Phys.*, vol. 23, pp. 1671-1684, 1996.

[15] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models,," *Int. J. Comput. Vision*, vol. 1, pp. 321-331, 1987.

[16] G. M. te Brake and N. Karssemeijer, "Segmentation of suspicious densities in digital mammograms," *Med. Phys.*, vol. 28, pp. 259-266, 2001.

[17] L. D. Cohen, "On active contour models and balloons," *CVGIP: Img. Underst.*, vol. 53, pp. 211-218, 1991.

[18] C. S. Poon and M. Braun, "Image segmentation by a deformable contour model incorporating region analysis," *Phys. Med. Biol.*, vol. 42, pp. 1833-1841, 1997.

[19] D. J. Williams and M. Shah, "A fast algorithm for active contours and curvature estimation," *CVGIP: Img. Underst.*, vol. 55, pp. 14-26, 1992.

[20] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pat. Anal. Mach. Intell.*, vol. 15, pp. 850-863, 1993.

[21] N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," *Med. Phys.*, vol. 26, pp. 1642-1654, 1999.

[22] L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications," *IEEE Trans. Med. Img.*, vol. 13, pp. 263-274, 1994.

[23] S. Mori, H. Nishida, and H. Yamada, *Optical Character Recognition.* New York: Wiley, 1999.

[24] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, G. M. M, and D. D. Adler, "Classification of masses on mammograms using a rubber-band straightening transform and feature analysis.," *Proc. SPIE Med. Img.*, vol. 2710, pp. 44-50, 1996.

[25] B. Sahiner, H. P. Chan, N. Petrick, G. M. M, and M. A. Helvie, "Characterization of masses on mammograms: Significance of the use of the rubber-band straightening transform," *Proc. SPIE Med. Img.*, vol. 3034, pp. 491-500, 1997.

[26] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Sys. Man. and Cybern.*, vol. SMC-3, pp. 610-621, 1973.

[27] M. M. Galloway, "Texture classification using gray level run lengths," *Comp. Graph. Img Proc.*, vol. 4, pp. 172-179, 1975.

[28] B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Med. Phys.*, pp. (in press), 2001.

[29] H.-P. Chan, N. Petrick, and B. Sahiner, "Chapter 6. Computer-aided breast cancer diagnosis," in *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*, A. Jain, A. Jain, S. Jain, and L. Jain, Eds. New Jersey: World Scientific, 2000, pp. 179-264.

[30] J. J. Koenderink and A. J. van Doorn, "Generic neighborhood operators," *IEEE Trans. Pat. Anal. Mach. Intell.*, vol. 14, pp. 597-605, 1992.

[31] N. R. Draper, *Applied regression analysis*. New York: Wiley, 1998.

[32] M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics*. Chicago, IL: SPSS Inc., 1993.

[33] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857-876, 1995.

[34] H. P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature space," *Med. Phys.*, vol. 25, pp. 2007-2019, 1998.

[35] N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," *Med. Phys.*, vol. 23, pp. 1685-1696, 1996.

[36] C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Stat. Med.*, vol. 17, pp. 1033-1053, 1998.

[37] C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance for differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging*, F. Deconinck, Ed. Martinus Nijhoff: The Hague, 1984, pp. 432-445.

[38] P. R. Detmer, G. Bashein, and R. W. Martin, "Matched filter identification of left-ventricular endocardial borders in transesophageal echocardiograms," *IEEE Trans. Med. Img.*, vol. 9, pp. 396-404, 1990.

Table Captions

Table I. The Hausdorff distance, the average minimum Euclidean distance, and the area overlap measure between R1 and R2; active contour segmentation (AC) and R1, and AC and R2.

Table II. The significance level p for the comparison of the inter-observer variation to the variation between computer-segmented and radiologist-segmented mass boundaries.

Figure Captions

Figure 1. The distribution of the malignancy rating of the masses in our data set, by an experienced radiologist. 1: Very likely benign, 10: Very likely malignant.

Figure 2. The distribution of the mass size for the 249 masses in our data set. Mass sizes were measured as the longest dimension of the mass by an experienced radiologist.

Figure 3. The distribution of the BI-RADS ratings for breast mass density for the 249 masses in our data set, by an experienced radiologist. 1=Almost entirely fat, 2=Scattered fibroglandular densities, 3=Heterogenerously dense, 4=Extremely dense breast.

Figure 4. The definition of the vertices and the normal direction used in the active contour model.

Figure 5. The graphical representation of the Hausdorff distance between contours A and B, which can be interpreted in this figure as the maximum of the minimum distances between any point on contour A and contour B. For an exact definition, please refer to the text. The area overlap measure is defined as the ratio of the cross-hatched area to the area of any hatched area.

Figure 6. The distribution of the Hausdorff distances between the segmentations by the active

contour (AC), R1, and R2.

Figure 7. The distribution of the average minimum Euclidean distances between the segmentations by the active contour (AC), R1, and R2.

Figure 8. The distribution of the area overlap measure between the segmentations by the active contour (AC), R1, and R2.

Figure 9. From top to bottom: The ROI, AC segmentation, R1 segmentation, and R2 segmentation for a (a) good AC segmentation (average AOM=0.92), (b) average AC segmentation (average AOM=0.73), and (c) poor AC segmentation (average AOM=0.50).

Figure 10. Per-view and per-case ROC curves for classifiers designed based on radiologist segmentation and computer segmentation.

Table I. The Hausdorff distance, the average minimum Euclidean distance (AMED), and the area overlap measure (AOM) between R1 and R2; active contour segmentation (AC) and R1, and AC and R2.

	Hausdorff distance (mm)		Average minimum Euclidean distance (mm)		Area overlap measure	
	Mean	Standard Dev.	Mean	Standard Dev.	Mean	Standard Dev.
R1 vs R2	2.5	1.9	0.85	0.69	0.76	0.13
AC vs. R1	2.6	1.5	0.90	0.50	0.74	0.11
AC vs. R2	2.8	1.5	0.93	0.58	0.74	0.13

Table II. The significance level p for the comparison of the inter-observer variation to the variation between computer-segmented and radiologist-segmented mass boundaries.

Similarity or Distance Measure	Significance	
	R1-R2 pairing and AC-R1 pairing	R1-R2 pairing and AC-R2 pairing
Hausdorff	0.52	0.07
AMED	0.34	0.09
AOM	0.10	0.02

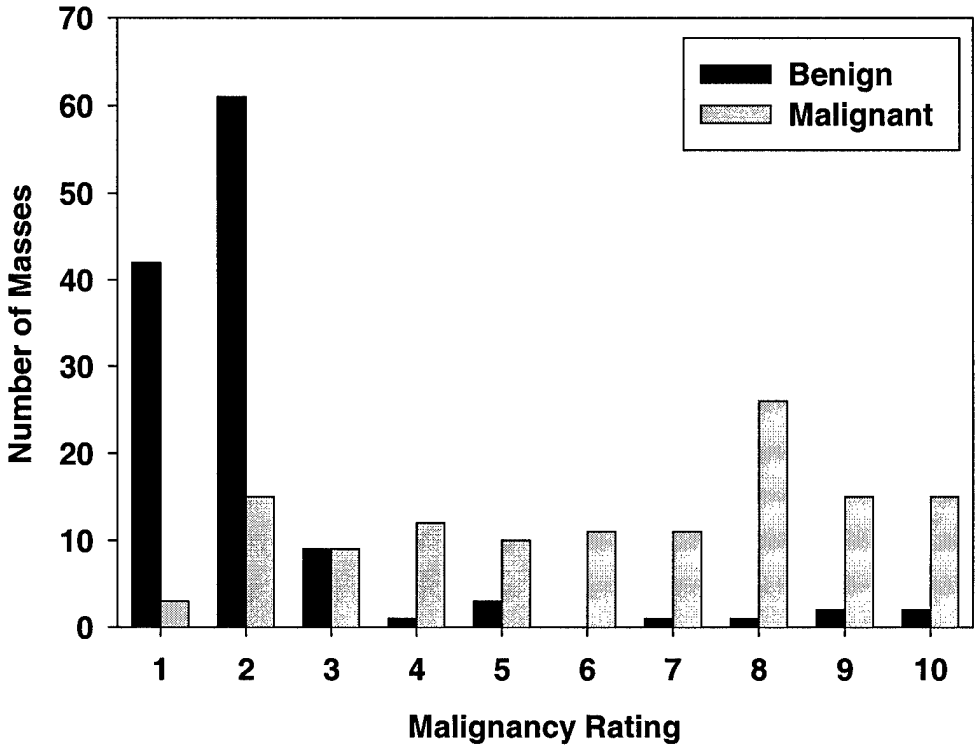


Figure 1. The distribution of the malignancy rating of the masses in our data set, by an experienced radiologist. 1: Very likely benign, 10: Very likely malignant.

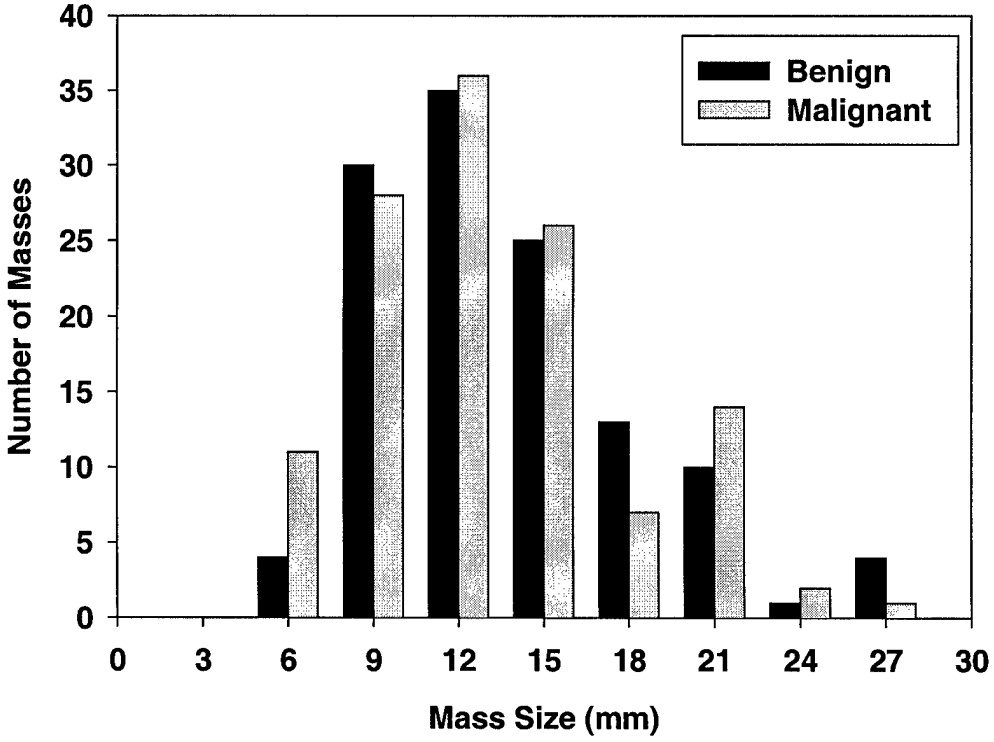


Figure 2. The distribution of the mass size for the 249 masses in our data set. Mass sizes were measured as the longest dimension of the mass by an experienced radiologist.

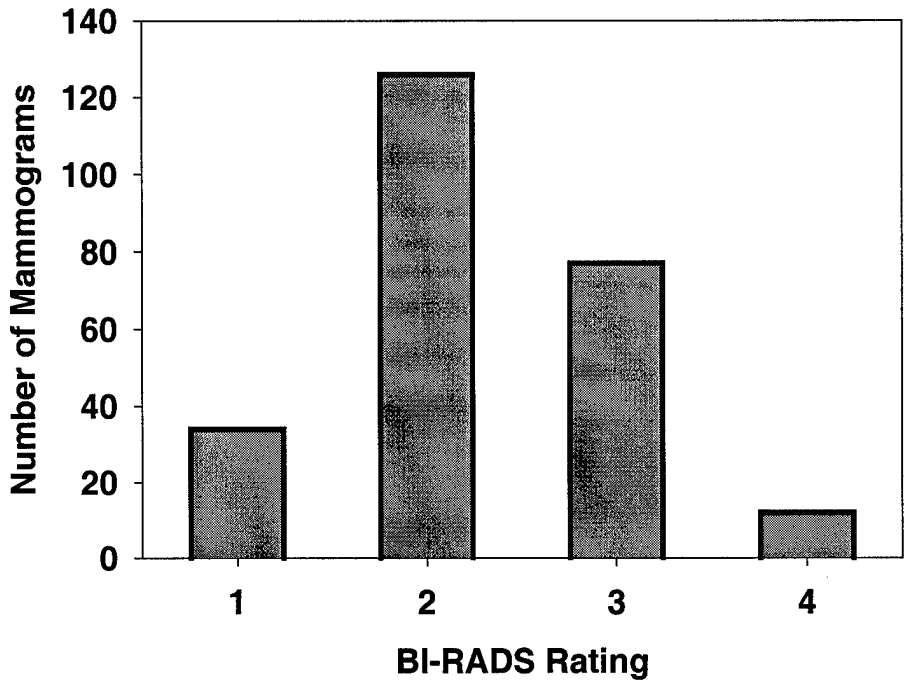


Figure 3. The distribution of the BI-RADS ratings for breast mass density for the 249 masses in our data set, by an experienced radiologist. 1=Almost entirely fat, 2=Scattered fibroglandular densities, 3=Heterogenerously dense, 4=Extremely dense breast.

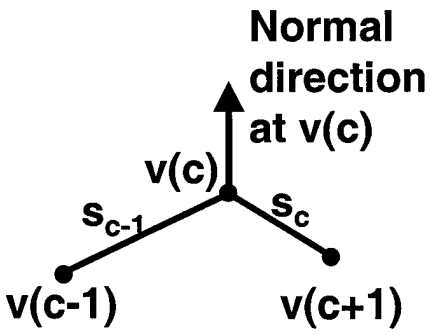


Figure 4. The definition of the vertices and the normal direction used in the active contour model.

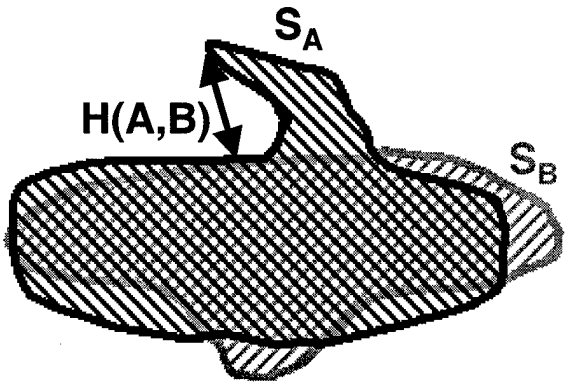


Figure 5. The graphical representation of the Hausdorff distance between contours A and B, which can be interpreted in this figure as the maximum of the minimum distances between any point on contour A and contour B. For an exact definition, please refer to the text. The area overlap measure is defined as the ratio of the cross-hatched area to the area of any hatched area.

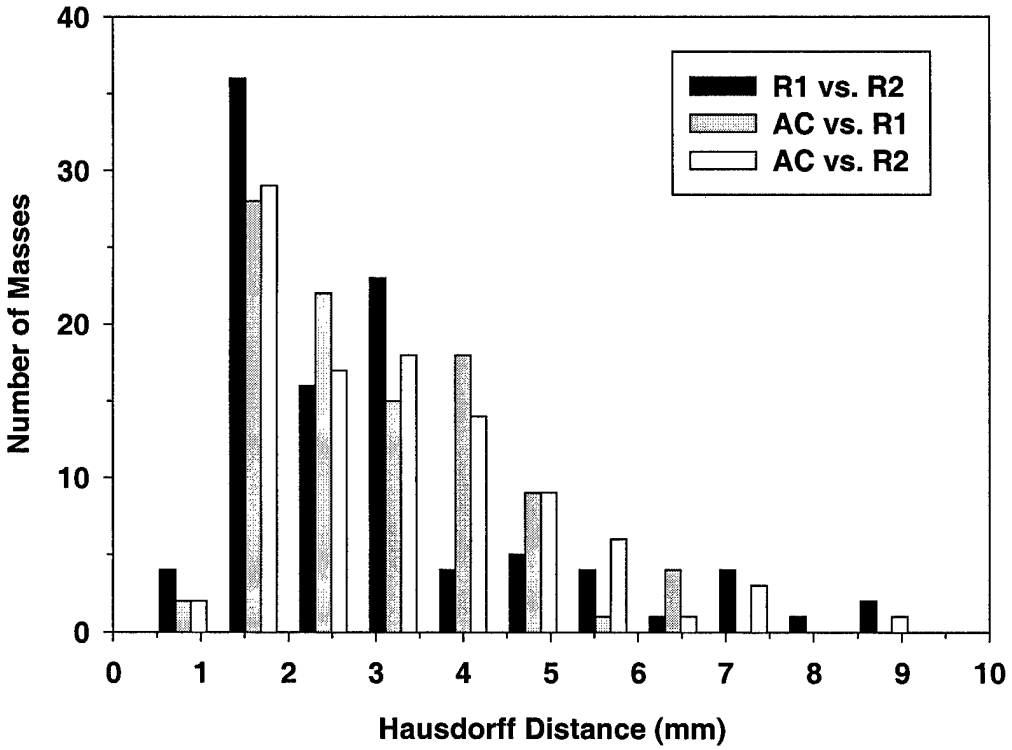


Figure 6. The distribution of the Hausdorff distances between the segmentations by the active contour (AC), R1, and R2.

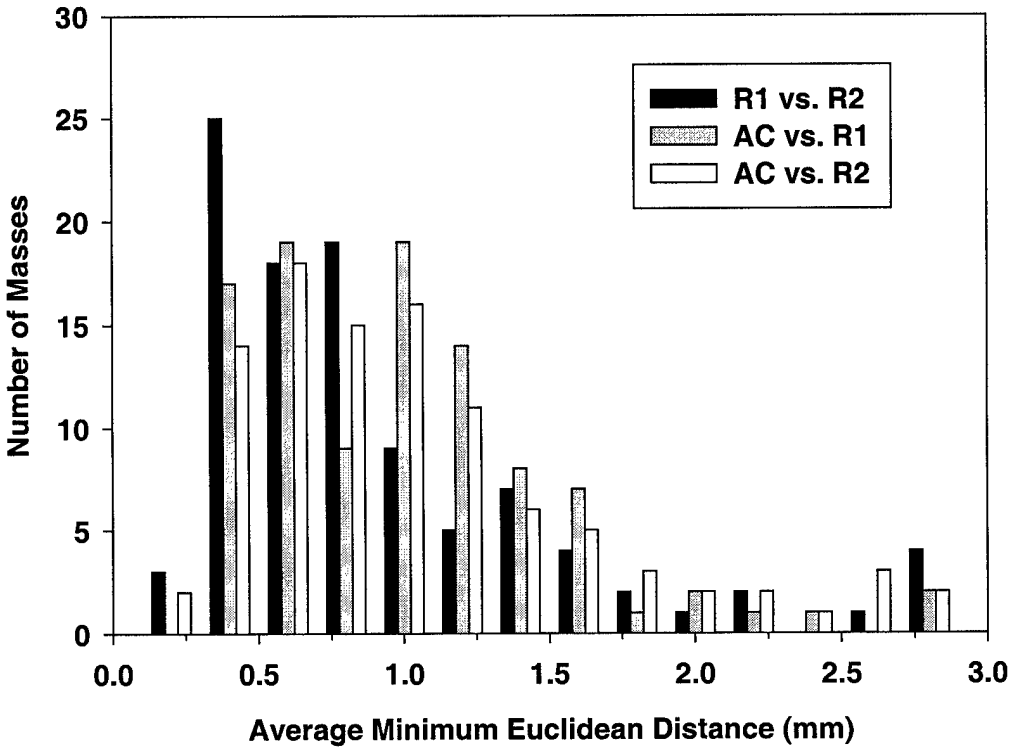


Figure 7. The distribution of the average minimum Euclidean distances between the segmentations by the active contour (AC), R1, and R2.

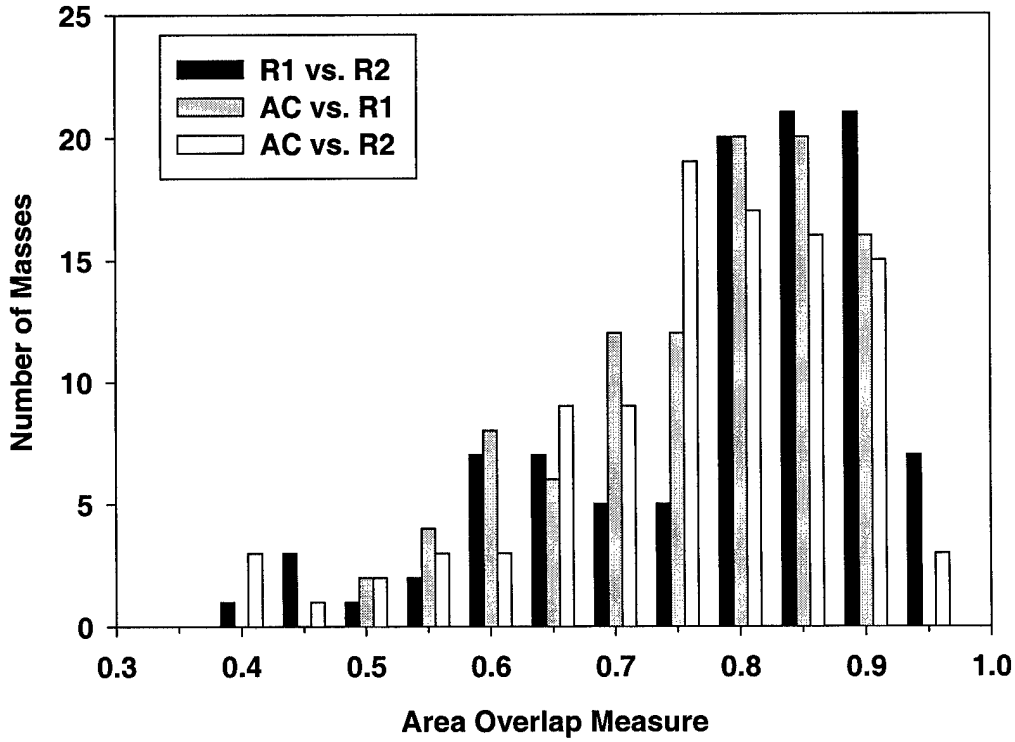


Figure 8. The distribution of the area overlap measure between the segmentations by the active contour (AC), R1, and R2.

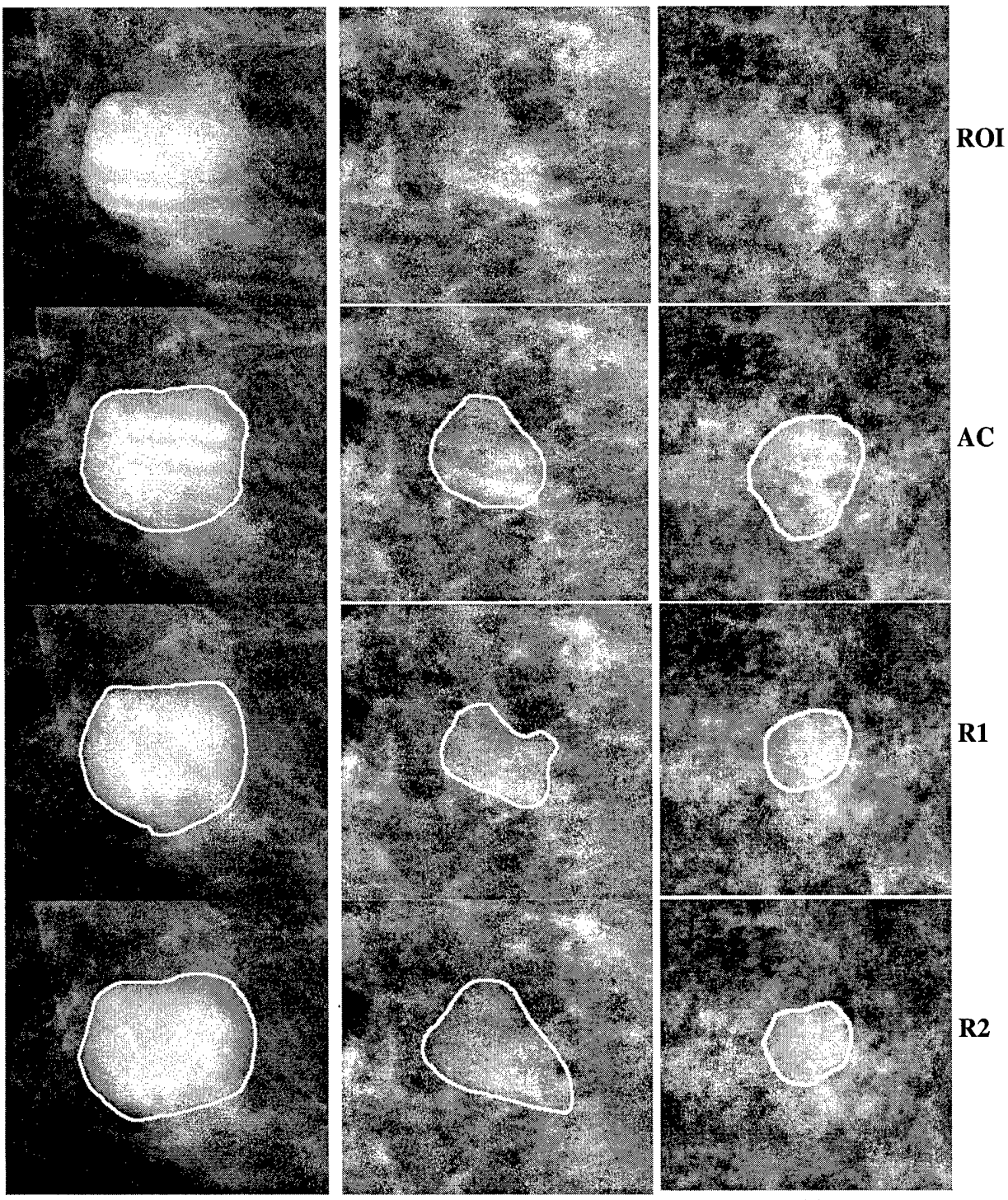


Figure 9. From top to bottom: The ROI, AC segmentation, R1 segmentation, and R2 segmentation for a (a) good AC segmentation (average AOM=0.92), (b) average AC segmentation (average AOM=0.73), and (c) poor AC segmentation (average AOM=0.50).

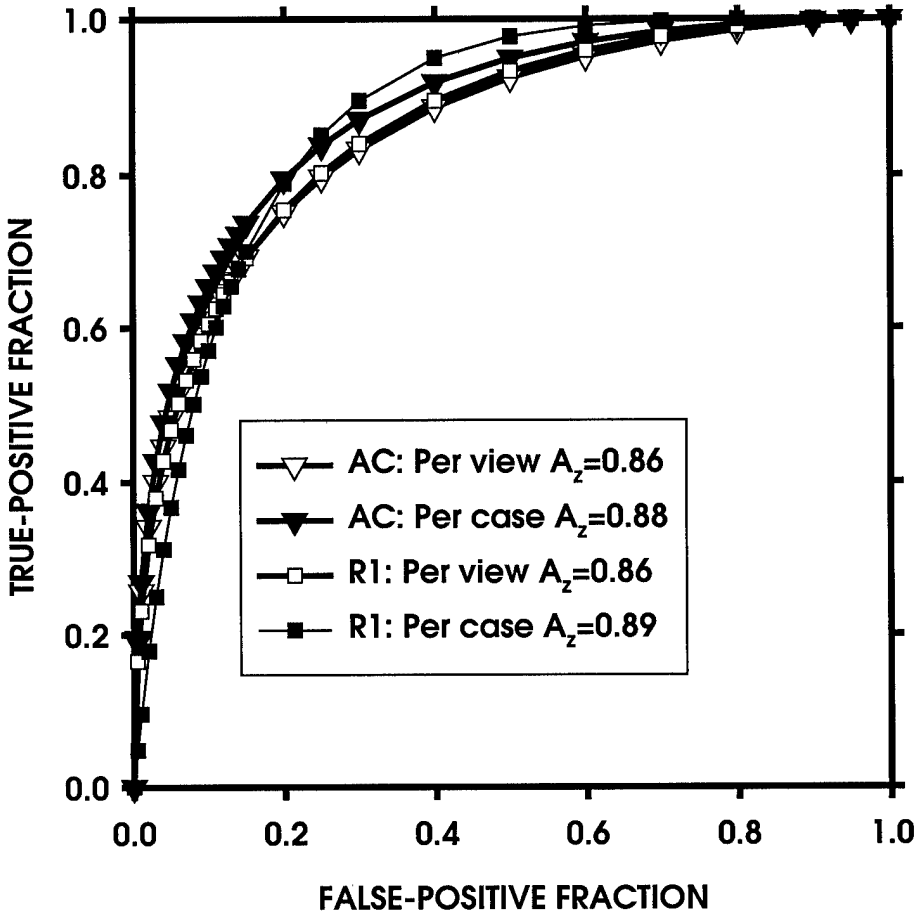


Figure 10. Per-view and per-case ROC curves for classifiers designed based on radiologist segmentation and computer segmentation.

Characterization of Masses on Mammograms: Significance of the Use of the Rubber-Band Straightening Transform

Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick,
Mitchell M. Goodsitt, and Mark A. Helvie
The University of Michigan
Department of Radiology
Ann Arbor, MI 48109-0030

ABSTRACT

The rubber-band straightening transform (RBST) was developed for characterization of mammographic masses as malignant or benign. The RBST maps a region surrounding a segmented mass on a mammogram onto the Cartesian plane. In this study, the effectiveness of texture features extracted from the RBST images was compared with the effectiveness of those extracted from the original images. Texture features were extracted from (i) a region of interest (ROI) centered at the mass; (ii) a 40-pixel-wide gray-scale region surrounding the perimeter of the mass; and (iii) the RBST image. Two types of texture features were extracted; spatial gray level dependence (SGLD) features and run-length statistics (RLS) features. Linear discriminant analysis and leave-one-case-out methods were used for classification in the individual or combined feature spaces. The classification accuracy was evaluated by Receiver Operating Characteristic (ROC) analysis and the area A_z under the ROC curve. CLABROC analysis was used to estimate the statistical significance of the difference between features extracted using the three different approaches. On a database of 255 ROIs containing biopsy-proven masses, the A_z value was 0.92 when combined SGLD and RLS features extracted from RBST images were used for classification. In comparison, the combined texture features extracted from the entire ROIs and the mass perimeter regions resulted in A_z values of 0.83 and 0.85, respectively. The improvement in A_z obtained by using RBST images was statistically significant ($p < 0.05$). Similar levels of significance were observed when the classification was performed in the SGLD feature space alone or the RLS feature space alone.

Keywords: Mammography, Computer-Aided Diagnosis, Masses, Classification, Texture Analysis, Discriminant Analysis, ROC Analysis.

1. INTRODUCTION

Computer-aided characterization of breast masses as malignant and benign has been an active area of research in recent years.[1-5] Important indicators in the discrimination of malignant and benign masses include characteristics of the mass borders and the region surrounding the mass.[6] The rubber band straightening transform (RBST) was introduced in order to transform the region surrounding a mammographic mass in such a way that texture orientations in the transformed image become more suitable for texture feature extraction using existing techniques.[7]

It was previously shown that texture features extracted from the RBST images are useful in discriminating malignant and benign masses.[7] However, the improvement in the classification accuracy due to the use of the

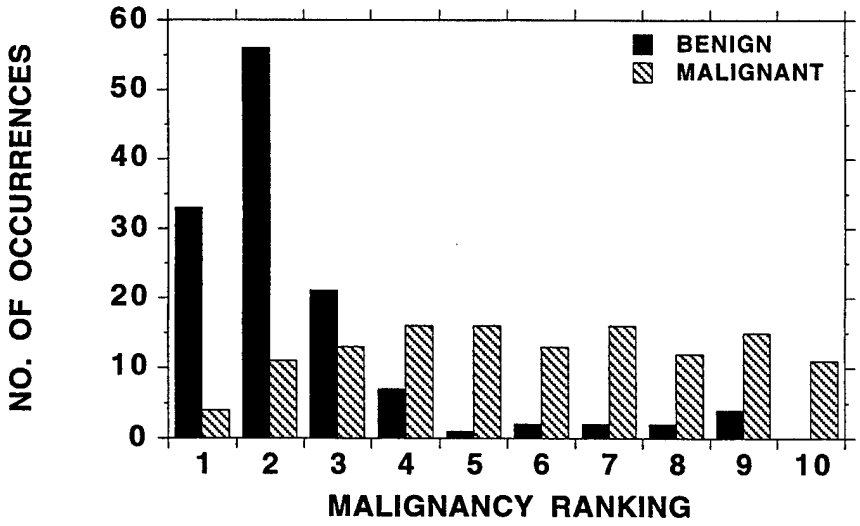


Figure 1: The distribution of the malignancy ranking of the masses in our dataset, as evaluated by an experienced radiologist.

RBST was not quantified. In this study, we compared the effectiveness of texture features extracted from the RBST images with the effectiveness of those extracted from the region surrounding the perimeter of the mass, or from a region of interest (ROI) containing the mass.

2. MATERIALS AND METHODS

2.1 Data Set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsy in the Department of Radiology at the University of Michigan. Our data set consisted of 255 mammograms, from 104 patients. There were 128 benign masses, of which 8 were spiculated, and 127 malignant masses, of which 62 were spiculated. The probability of malignancy of the biopsied mass on each mammogram was ranked by an experienced breast radiologist on a scale of 1 to 10 based on its mammographic appearance. A ranking of 1 corresponded to the masses with the most benign appearance, and a ranking of 10 corresponded to the masses with the most malignant appearance. The distribution of the malignancy ranking of the masses is shown in Fig. 1. The true nature of the mass was determined by biopsy and histologic analysis.

The mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel size of $100\mu m \times 100\mu m$ and 4096 gray levels. The location of the biopsied mass was identified by an experienced radiologist, and a variable-size ROI centered around the identified mass was extracted from the digitized mammogram.

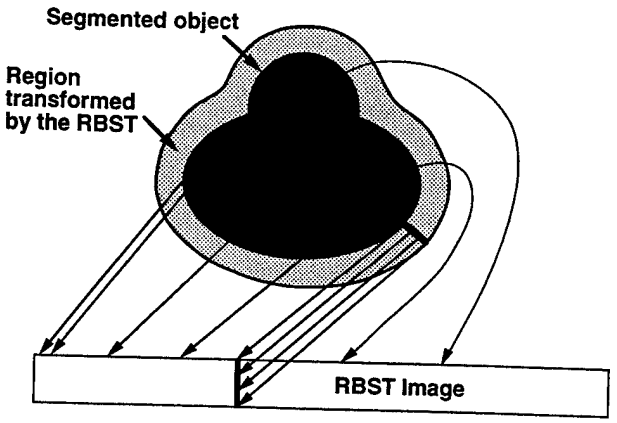


Figure 2: The rubber band straightening transform (RBST). The pixels along the object boundary are mapped to the first row of the RBST image. Pixels on a normal line to the object are mapped to a column of the RBST image.

2.2 The RBST image

The RBST maps a region surrounding a segmented mass on a mammogram onto the Cartesian plane. This operation is depicted in Fig. 2. The block diagram for the computation of an RBST image is shown in Fig. 3.

The first step in the computation of the RBST, as well as any other processing (*e.g.*, feature extraction or segmentation) on an ROI in this paper was background correction. A background image was estimated using a linear combination of a band of pixels along the perimeter of the ROI.[8] This background image was then subtracted from the original image, thus reducing the non-uniform background within the ROI.

After background correction, the ROI was segmented into a mass object and background tissue. The segmentation was based on a modification of the K-means clustering algorithm, described in detail elsewhere.[9] The result of the clustering algorithm was a detected object, or a small number of detected objects in the ROI. If more than one object was detected, the largest connected object among all objects was selected. The selected object was then filled, grown in a local neighborhood, and eroded and dilated with morphological operators. The details of the segmentation algorithm have been described previously.[7]

Three main steps of the RBST are edge enumeration, computation of normals, and interpolation, which are explained next. The edge enumeration algorithm assigned a number to each border pixel of the segmented mass. The algorithm was designed so that neighboring pixels were assigned consecutive numbers. The computation of the normal direction to the object was based entirely on the result of the edge enumeration algorithm. Consider a pixel that was assigned location i by the edge enumeration algorithm. To find the normal line $L(i)$ at this pixel location in the ROI, the two pixels that were assigned $i+K$ and $i-K$ by the edge enumeration algorithm were located in the ROI. $L(i)$ was found as the normal to the line joining these two pixels. It was empirically found that $K = 12$ resulted in acceptable normal direction computation for the masses in our database.

In the interpolation step, the value of the pixel in row j, column i of the RBST image was found as follows. Let $p(i,j)$ denote the point on $L(i)$ which has distance j from pixel i. The two closest pixels to $p(i,j)$ were identified, and the corresponding pixel value of the RBST image was defined as the distance-weighted average of these two pixel values. In this paper, we used a 40-pixel-wide region of the ROI surrounding the object to determine the RBST image. An example of an original ROI, segmented mass object, and the RBST image are given in Fig. 4.

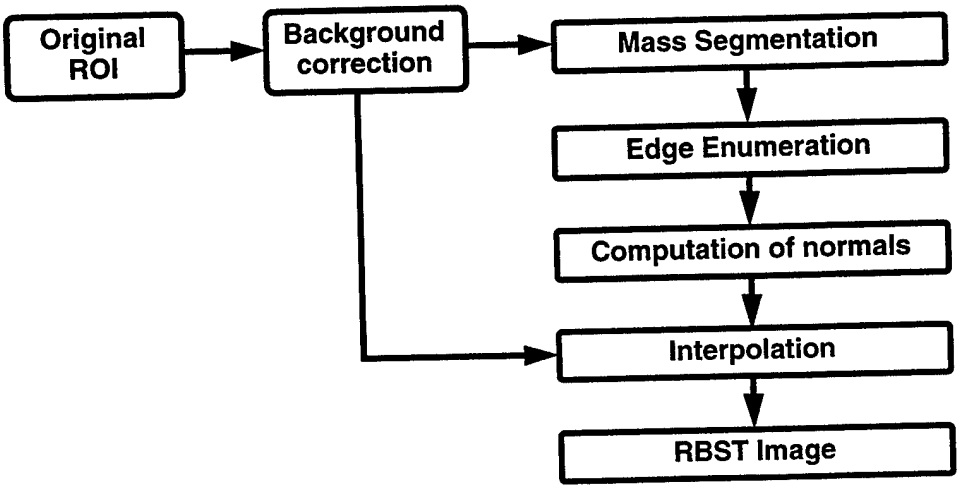


Figure 3: Block diagram of the computation of an RBST image.

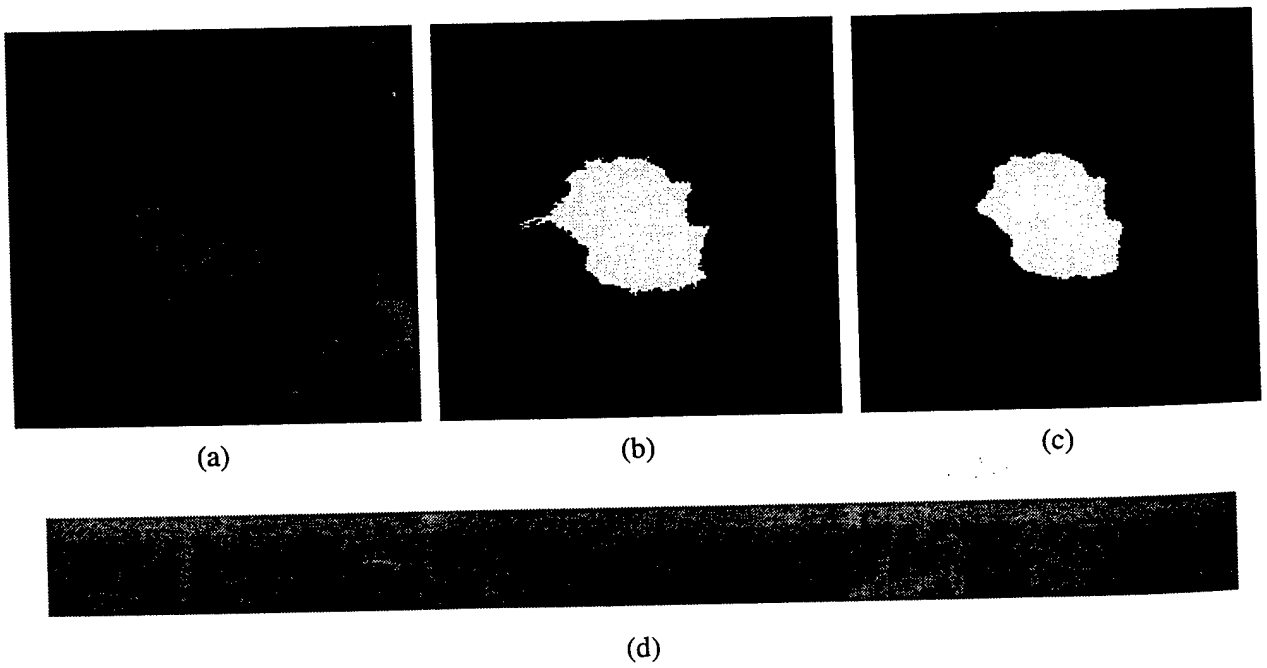


Figure 4: (a) Original image, (b) segmented mass before erosion and dilation, (c) segmented mass after erosion and dilation, (d) RBST image.

2.3 Texture Features

The texture features used in this study were calculated from the run-length statistics (RLS) matrices[10] and the spatial gray-level dependence (SGLD) matrices.[11,12] The construction of these matrices and the significance of the features extracted from them are described in the literature.[10,13] Each of these matrices was constructed separately using three different representations of the same mass. These representations were: $R1$, the entire ROI containing the mass; $R2$, a 40-pixel-wide region of the ROI surrounding the segmented mass; and $R3$, the RBST image.

The RLS matrices were constructed using the vertical and horizontal gradient magnitudes of each image representation. Five features (short runs emphasis, long runs emphasis, gray level nonuniformity, run length nonuniformity, and run percentage) were extracted from each matrix in two directions ($\theta = 0°$ and $\theta = 90°$). Therefore, the total number of RLS features extracted for each image representation was 20.

The SGLD matrices were constructed using the gray-level values of each image representation. Eight texture features (correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy and inertia) were extracted at ten different pixel pair distances (d=1, 2, 3, 4, 6, 8, 10, 12, 16, and 20) and in four directions ($\theta = 0°$, $\theta = 45°$, $\theta = 90°$, and $\theta = 135°$). Therefore, the total number of SGLD features extracted for each image representation was 320.

We thus had an RLS feature space with 20 texture features, an SGLD feature space with 320 texture features, and a combined texture space, which contains 340 texture features. The effectiveness of each of these three feature spaces in the classification of mammographic masses was compared for the three image representations, $R1$, $R2$, and $R3$.

2.4 Classification

Linear discriminant analysis[14,15] with stepwise feature selection was used to classify malignant and benign masses based on the extracted texture features. The selection criterion in the stepwise algorithm was Wilks' lambda, defined as the ratio of the within-group sum of squares to the total sum of squares. At each step of the algorithm, the significance of the change in the Wilks' lambda when a new feature was entered into the selected feature pool was compared to a threshold F_{in}. The feature with the highest significance was added to the selected feature pool only if the significance was higher than F_{in}. Likewise, the significance of the change in the Wilks' lambda when a selected feature was removed from the feature pool was compared to a threshold F_{out}. The feature with the highest significance was removed from the selected feature pool only if the significance was higher than F_{out}. Since the optimal values of F_{in} and F_{out} are not known *a priori*, we varied these thresholds to obtain the best test performance for a given image representation and a given feature space.

A leave-one-case-out method was used to train and test the classifier.[7] The discriminant scores were used as the decision variable in the LABROC1 program,[16] which provided the ROC curve based on maximum likelihood estimation. The area A_z under the ROC curve was used as an index of classification accuracy. The CLABROC program[17] was used to test the difference of the classification accuracy with features extracted from different image representations.

3. RESULTS

The test results with the highest A_z value obtained using the SGLD feature space are summarized in Table 1, and the corresponding ROC curves are plotted in Fig. 5. Each line in Table 1 was obtained by varying the F_{in} and the F_{out} values in stepwise feature selection to "optimize" the test A_z value for a given image representation.

The difference between classification results using $R1$ and $R3$ was statistically significant ($p < 0.05$). The difference between $R2$ and $R3$ did not achieve statistical significance.

The test results with the highest A_z value obtained using the RLS feature space are summarized in Table 2, and the corresponding ROC curves are plotted in Fig. 6. The difference between classification results using $R1$ and $R3$, as well as $R2$ and $R3$ were statistically significant ($p < 0.05$).

Table 1: Classifier performance with SGLD texture features

Representation	F_{in}	F_{out}	Num. of Features	Training A_z	Test A_z
$R1$	1.0	0.8	37	0.91 ± 0.02	0.81 ± 0.03
$R2$	1.6	1.4	27	0.90 ± 0.02	0.84 ± 0.02
$R3$	1.1	0.6	28	0.94 ± 0.01	0.89 ± 0.02

Table 2: Classifier performance with RLS texture features

Representation	F_{in}	F_{out}	Num. of Features	Training A_z	Test A_z
$R1$	0.6	0.4	5	0.69 ± 0.03	0.65 ± 0.03
$R2$	1.6	1.4	2	0.67 ± 0.03	0.65 ± 0.03
$R3$	2.4	2.2	6	0.81 ± 0.03	0.78 ± 0.03

Table 3: Classifier performance with combined texture features

Representation	F_{in}	F_{out}	Num. of Features	Training A_z	Test A_z
$R1$	1.2	1.0	33	0.91 ± 0.02	0.83 ± 0.03
$R2$	1.2	1.0	26	0.91 ± 0.02	0.85 ± 0.02
$R3$	1.4	1.2	41	0.97 ± 0.01	0.92 ± 0.02

Finally, the corresponding table and the ROC curves obtained using the combined feature space are shown in Table 3 and Fig. 7. Again, the difference between classification results using $R1$ and $R3$, as well as $R2$ and $R3$ were statistically significant ($p < 0.05$). The distribution of the discriminant scores of the classifier designed with the combined features extracted from RBST images is shown in Fig. 8.

A comparison of Figs. 1 and 8 reveals that the distribution of the computer classifier scores and the malignancy ranking of the radiologist are different. The radiologist's ranking of the malignant masses distributes across the entire x-axis, and that of the benign masses concentrates mainly at low (1 to 2) rankings with a long tail extending to a rating of 9. The computer scores for both the malignant and benign masses seem to be closer to having a Gaussian distribution. The radiologist's ratings (Fig. 1) seem to agree with the fact that all of the masses in our database were biopsied; if no false negatives are desired, then all suspicion levels have to be biopsied. However, using the computer scores in Fig. 8, if a mass is considered to be suspicious when its discriminant score is greater than 4, more than 30% of the benign masses can be correctly classified without any false negatives.

4. CONCLUSION

In this paper, we investigated the effectiveness of the texture features extracted from the RBST images for discrimination of malignant and benign masses. With the best combination of SGLD and RLS features, the test A_z value for our database of 255 mammograms was 0.92. With the use of the correct threshold, more than 30% of benign masses could be correctly classified with no missed malignant masses.

Since the RBST image is obtained from the 40-pixel-wide band surrounding the segmented mass, we compared the effectiveness of the texture features extracted from the RBST images to the effectiveness of those extracted from a 40-pixel-wide band surrounding the mass. We found that in two of the three features spaces that we investigated (RLS and combined feature spaces), the features extracted from the RBST images were significantly more effective. Compared to texture features extracted from the entire ROI, those extracted from the RBST images were significantly more effective in all three feature spaces. Future work on characterization of masses using RBST includes the development of better mass segmentation methods, and the investigation of the combined information from multiple views.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award (B. S.) from the USAMRMC (DAMD 17-96-1-6012) and a USPHS Grant CA 48129. The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC1 and CLABROC programs.

REFERENCES

1. J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," *IEEE Transactions on Medical Imaging*, vol. 12, pp. 664–669, 1993.

2. S. Pohlman, K. A. Powell, N. A. Obuchowski, W. A. Chilcote, and S. G-Broniatowski, "Quantitative classification of breast tumors in digitzed mammograms," *Medical Physics*, vol. 23, pp. 1337–1345, 1996.

3. Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," *Medical Physics*, vol. 22, pp. 1569–1579, 1995.

4. R. M. Rangayyan, N. El-Faramawy, J. E. L. Desautels, and O. A. Alim, "Discrimination between benign and malignant breast tumors using a region-based measure of edge profile acutance," in *Digital Mammography '96* (K. D. et. al., ed.), pp. 213–218, Amsterdam: Elsevier, 1996.

5. B. Sahiner, H.-P. Chan, N. Petrick, , M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis." Accepted for publication in Medical Physics, 1997.

6. L. Tabar and P. B. Dean, *Teaching Atlas of Mammography.* New York: Thieme, 1985.

7. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of masses on mammograms using rubber-band straightening transform and feature analysis," in *Proceedings of SPIE Medical Imaging: Image Processing*, vol. 2710, (Newport Beach, CA), pp. 44–50, Feb. 1995.

8. B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," *IEEE Transactions on Medical Imaging*, vol. 15, pp. 598–610, 1996.

9. B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: An artificial neural network with morphological features," in *Proceedings of the World Congress on Neural Networks*, (Washington, DC), pp. 876–879, July 1995.

10. M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, pp. 172–179, 1975.

11. H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," *Physics of Medicine and Biology*, vol. 40, pp. 857–876, 1995.

12. D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Medical Physics*, vol. 22, pp. 1501–1513, 1995.

13. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, pp. 610–621, 1973.

14. P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner Press, 1975.

15. M. J. Norusis, *SPSS Professional Statistics 6.1*. Chicago: SPSS Inc., 1993.

16. C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously distributed test results." presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim, CA, Aug. 1990.

17. C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance of differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging: Proceedings of the 8th conference* (F. Deconinck, ed.), pp. 432–445, Boston, Brussels: Martinus Nijhoff, 1984.

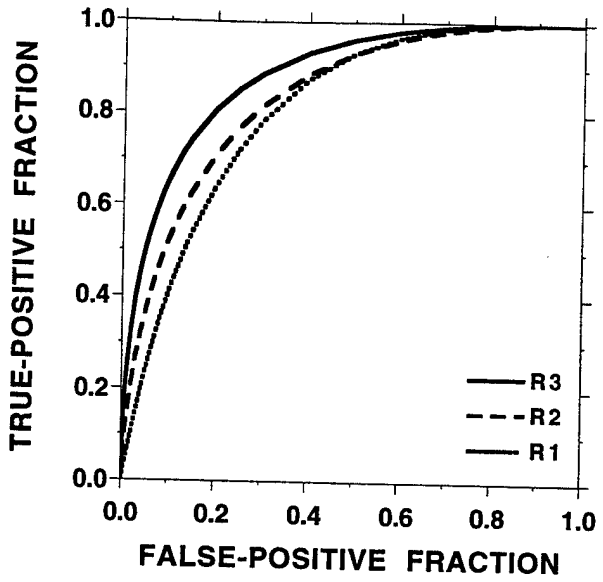


Figure 5: ROC curves for the three image representations using the SGLD feature space.

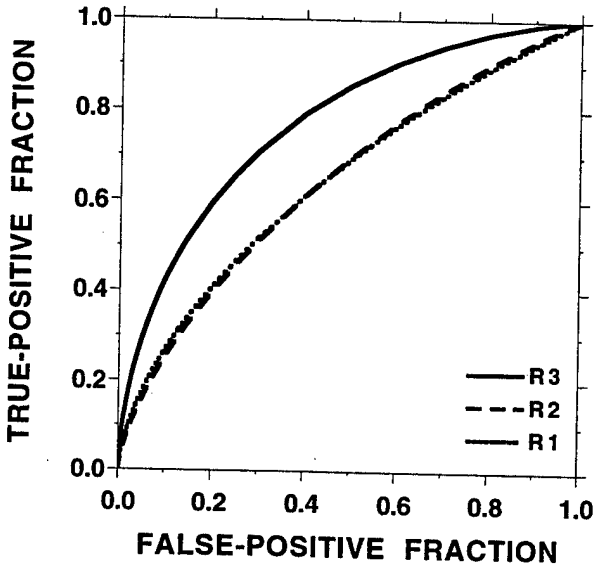


Figure 6: ROC curves for the three image representations using the RLS feature space.

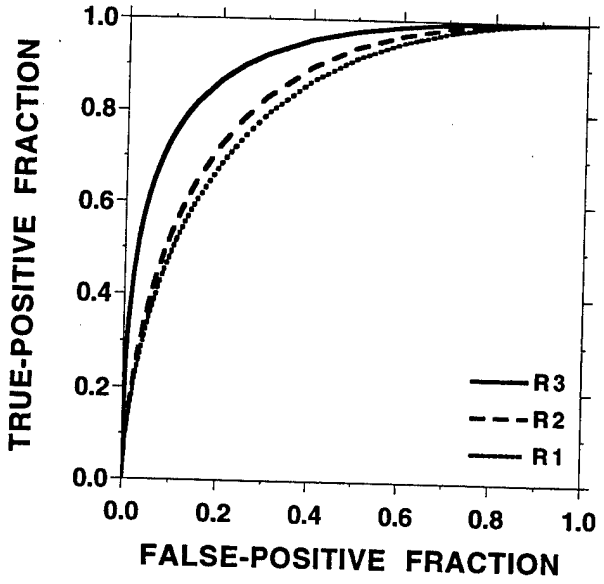


Figure 7: ROC curves for the three image representations using the combined feature space.

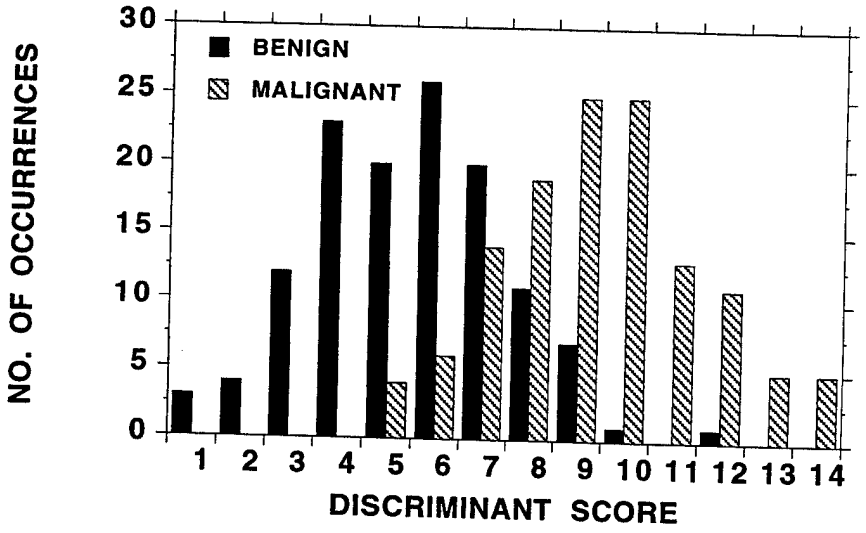


Figure 8: The distribution of the discriminant scores of the classifier designed with the combined features extracted from RBST images.

Neural Network Design for Optimization of the Partial Area Under the Receiver Operating Characteristic Curve

Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, S. Sanjay Gopal, and Mitchell M. Goodsitt

The University of Michigan
Department of Radiology
UH B1D403
Ann Arbor, MI 48109-0030
e-mail: berki@umich.edu

Abstract

A new backpropagation training algorithm was developed for the maximization of the area under the Receiver Operating Characteristic (ROC) curve between two user-specified true-positive fraction thresholds. The algorithm was used to design a neural network classifier with high specificity at the high-sensitivity region of the ROC curve, which is of particular interest for computer-aided diagnosis applications. The effectiveness of the algorithm was demonstrated with a simulation study.

1. Introduction

Backpropagation neural networks (BPNs) have long been used for two-group or multi-group classification problems [1]. In this paper, we are interested in two-group classification, and the BPN is assumed to have a single output node. In a conventional BPN used for a two-group classification task, the error of the network for a particular training sample is defined as the squared difference between the actual neural network output and the desired output for the sample. The network is trained to minimize the average error for the entire training population.

In many medical applications, the cost of missing a positive (e.g., diseased) case and the cost of misclassifying a negative (e.g., normal) case are not the same. The decision threshold therefore cannot be determined without a well-designed cost-benefit analysis. Receiver Operating Characteristic (ROC) analysis is a commonly-used methodology for representing the tradeoff between the true-positive fraction (TPF) and the false-positive fraction (FPF) in a two-group classification task. The area A_z under the ROC curve, or the area $A(T_0, T_1)$ under the ROC curve between two TPF thresholds T_0 and T_1, may be used to measure the classification accuracy. In several applications, such as computer-aided diagnosis (CAD), the latter index with $T_1 = 1$ (area under the ROC curve above a true-positive fraction T_0) may be considered to be a more desirable measure than A_z [2]. In a previous study [3], we investigated the maximization of $A(T_0, 1)$ using a genetic algorithm for feature selection.

In this paper, we develop an algorithm to maximize $A(T_0, T_1)$ for a BPN used for solving a two-group classification task. The maximization of $A(T_0, 1)$ is equivalent to the design of a classifier with high specificity at the high-sensitivity portion of the ROC curve, which has particular importance in applications such as CAD. The maximization of A_z is obtained as a special case for which $T_0 = 0$ and $T_1 = 1$. The design tradeoffs and the classification accuracy of the designed network are illustrated with a simulation study.

2. Partial area under the ROC curve

Most ROC curves are generated under the assumption that the decision variables for the two classes are either binormal, or can be so transformed by a monotonic transformation. Let the decision variables for the positive and negative classes be denoted by S and N, respectively, and assume that S and N are independent and normally distributed for the two classes, so that $S \sim \mathcal{N}(\mu_s, \sigma_s)$ and $N \sim \mathcal{N}(\mu_n, \sigma_n)$. The area $A(T_0, T_1)$ under the ROC curve between two TPF thresholds T_0

0-7803-4122-8/97 $10.00

and T_1 can be expressed as

$$A(T_0,T_1) = (T_1 - T_0) - \int_{T_0}^{T_1} FPF(TPF)dTPF. \quad (1)$$

Under the assumption that the decision variables are normal, the false-positive and true-positive rates are related by [4]

$$\Phi^{-1}(TPF)\sigma_s - \mu_s = \Phi^{-1}(FPF)\sigma_n - \mu_n, \quad (2)$$

where $\Phi(x)$ is the cumulative Gaussian distribution function,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du, \quad (3)$$

and Φ^{-1} is its inverse.

In an analogous manner to[5], $A(T_0,T_1)$ can be expressed using (1) and (2) as

$$A(T_0,T_1) = (T_1 - T_0) - \int_{c_0}^{c_1} \Phi\left(\frac{\sigma_s u + d_m}{\sigma_n}\right) \phi(u) du, \quad (4)$$

where

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad (5)$$

$$d_m = \mu_n - \mu_s, \quad (6)$$

and

$$c_i = \Phi^{-1}(T_i), \quad i = 0, 1. \quad (7)$$

3. BPN training with partial ROC area criterion

The training of a BPN via the gradient-descent rule involves the computation of the partial derivatives of the network error with respect to the network weights. In this paper, the network error to be minimized is defined as

$$E = -A(T_0,T_1), \quad (8)$$

so that

$$\frac{\partial E}{\partial w(i,j)} = -\frac{\partial A(T_0,T_1)}{\partial w(i,j)}, \quad (9)$$

where $w(i,j)$ is the weight between nodes i and j in the neural network. Assuming that there are P training samples, this partial derivative can be written as

$$\frac{\partial E}{\partial w(i,j)} = -\sum_{p=1}^{P} \frac{\partial A(T_0,T_1)}{\partial O^{(p)}} \frac{\partial O^{(p)}}{\partial w(i,j)}, \quad (10)$$

where $O^{(p)}$ denotes the output of the neural network corresponding to sample p.

The sum in (10) contains two terms. The first term is related to the partial derivative of $A(T_0,T_1)$ with respect to the neural network output, and the second term is related to the partial derivative of the neural network output with respect to the neural network weights. This second term is commonly encountered in all backpropagation networks, and can be computed using the standard backpropagation rule.

In this paper, we propose to use the partial derivatives of (4) to approximate the first term. Notice that (4) is derived under the assumption that the decision variables are binormal for the two classes. Although this assumption may not be satisfied for a well-trained neural network, we have chosen to use the binormality framework because: (i) it has been shown to be a robust approximation in fitting ROC curves to discrete-rating data [6], and the computation of the ROC curve for continuous data (LABROC) involves categorization to discrete-rating data [4]; and (ii) it provides us with closed-form expressions that are relatively simple to evaluate.

Under the binormality assumption, the partial derivative of $A(T_0,T_1)$ with respect to the neural network output can be analytically computed by finding the partial derivatives with respect to σ_s, σ_n, and d_m. It follows that the derivatives of $A(T_0,T_1)$ with respect to $w(i,j)$ are available. Steepest descent rule with the computed partial derivatives results in a BPN that learns to maximize $A(T_0,T_1)$.

4. Simulation Study and Discussion

To test our new BPN training algorithm, we used a randomly-generated six-dimensional Gaussian dataset with means $\mathbf{M_p}$ and $\mathbf{M_n}$, and diagonal covariance matrices $\mathbf{\Lambda_p}$ and $\mathbf{\Lambda_n}$ for the positive and negative cases, respectively. The means and covariances were chosen as $\mathbf{M_p} = 0.5 * [1,1,1,1,1,1]^T$, $\mathbf{M_n} = -0.5 * [1,1,1,1,1,1]^T$, $\mathbf{\Lambda_p} = diag(3,3,3,1,1,1)$, and $\mathbf{\Lambda_n} = diag(1,1,1,3,3,3)$. The total number of cases for each class was 300. Two hundred randomly-drawn samples from each class were used for training the BPN, and the remaining one hundred samples were used for testing.

The BPN was trained in batch mode, *i.e.*, the weight updates were performed at the end of each training epoch. This meant that the neural network weights and the output due to a given training sample did not change within an epoch, and therefore the shape of the ROC curve was fixed within an epoch. As a result, most of the computations involved in the calculation of the partial derivatives of $A(T_0,T_1)$ needed to be performed only once during each epoch. This lead to a fast implementation of the new training algorithm.

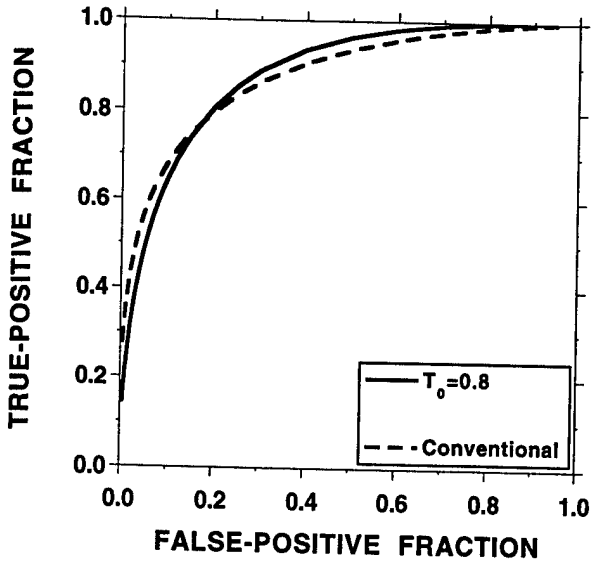


Figure 1. ROC curves for neural networks with 2 hidden layer nodes

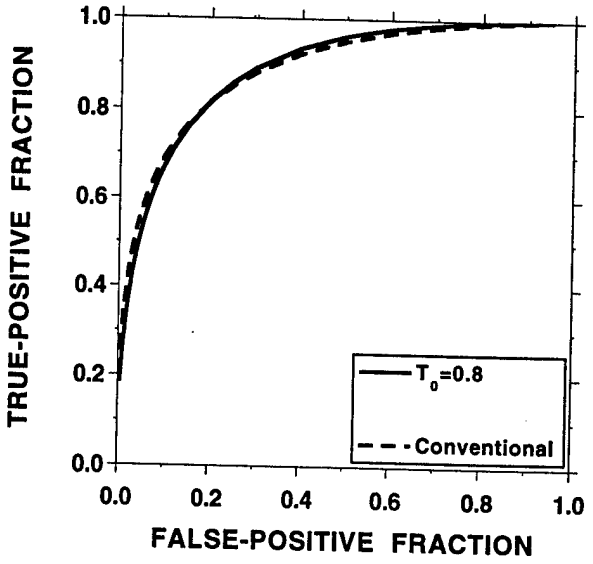


Figure 3. ROC curves for neural networks with 5 hidden layer nodes

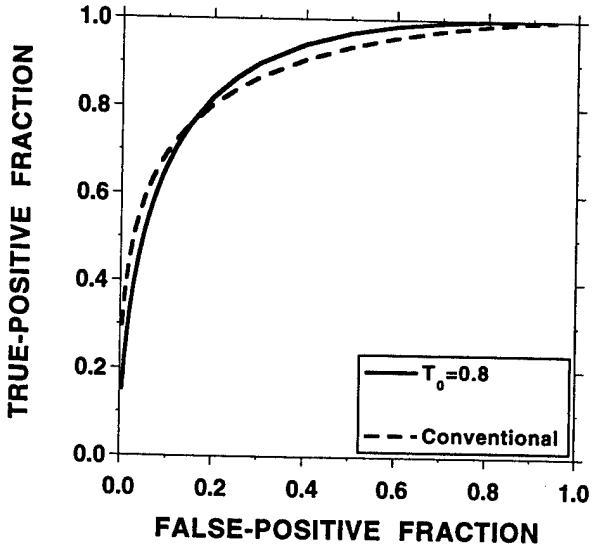


Figure 2. ROC curves for neural networks with 3 hidden layer nodes

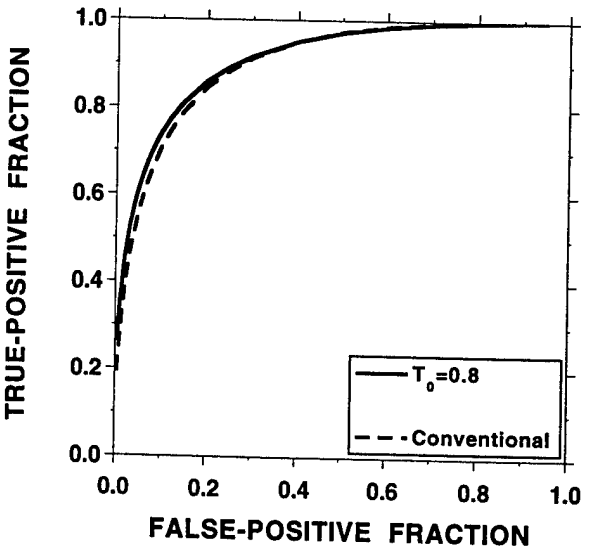


Figure 4. ROC curves for neural networks with 9 hidden layer nodes

Since our goal in this study was to obtain a high sensitivity classifier, the upper limit of the partial area T_1 was fixed at $T_1 = 1.0$. We studied the effect of the number of hidden-layer nodes on the conventional and the high sensitivity training algorithms by varying the number of hidden-layer nodes between 2 and 9.

For both training algorithms, the neural network was trained starting from the same initial condition. After 5000 training iterations, test outputs were obtained by applying the test samples to the trained networks. ROC curves were fitted using the LABROC program by Metz *et. al* [4], and ROC curve parameters a and b were estimated. This procedure was repeated for ten different initializations of the weights based on a random number generator. A final averaged ROC curve was obtained by averaging a and b parameters over different initializations.

Figs. 1-4 show the test ROC curves for the new training algorithm with $T_0 = 0.8$ and the conventional training algorithm for different number of hidden-layer nodes. For small number of hidden-layer nodes, the new training algorithm achieves the goal of decreasing the false-positives for a TPF of 0.8 and above, as observed from Figs. 1 and 2. To compensate, the FPF at small TPFs are increased. However, this increase would not be important in many medical applications such as in CAD, where a TPF of less than 0.8 would probably be unacceptable, and hence the false positive rates below this threshold would be of no consequence.

Figs. 3 and 4 show that the difference between the two training algorithms diminishes as the number of hidden-layer nodes are increased. This may be due to the fact that when the neural network has enough training samples and a large number of hidden-layer nodes, the performance of the neural network is close to the performance of an optimal classifier, and there is less possibility for improvement. In many practical situations, one would have a limited number of training samples, and using a large number of hidden-layer nodes could result in an overtrained classifier with poor performance when tested on independent samples.

5. Conclusion and future work

We have developed a new BPN training algorithm to maximize the area under the ROC curve between two given TPF thresholds. The algorithm can be used to design a classifier with a high specificity in the high-sensitivity region of the ROC curve. The simulation study demonstrates the usefulness of such a training strategy when the number of hidden layer nodes is small, which represents desired network architectures to prevent overtraining if a small number of training samples is available, such as in medical applications.

We are currently investigating the effect of the feature distributions on the effectiveness of the new algorithm, and its convergence properties.

Acknowledgments

This work is supported by a Career Development Award (B. S.) from the USAMRMC (DAMD 17-96-1-6012), a USPHS Grant CA 48129, and a USAMRMC grant CDAMD 17-96-1-6254. The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment or product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC1 program.

References

[1] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: An artificial neural network with morphological features," in *Proceedings of the World Congress on Neural Networks*, Washington, DC, July 1995, pp. 876–879.

[2] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," *Radiology*, vol. 201, pp. 745–750, 1996.

[3] B. Sahiner, H.-P. Chan, N. Petrick, , M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of malignant and benign breast masses: Development of a high-sensitivity classifier using a generic algorithm," *Radiology*, vol. 201(P), p. 257, 1996.

[4] C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously distributed test results," presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim, CA, Aug. 1990.

[5] D. K. McClish, "Analyzing a portion of the ROC curve," *Medical Decision Making*, vol. 9, pp. 190–195, 1989.

[6] J. A. Hanley, "The robustness of the "binormal" assumptions used in fitting ROC curves," *Medical Decision Making*, vol. 8, pp. 197–203, 1988.

Neural Network Based Segmentation Using *a priori* Image Models

S. Sanjay Gopal, Berkman Sahiner, Heang-Ping Chan, and Nicholas Petrick
The University of Michigan
Department of Radiology
UH B1D403
Ann Arbor, MI 48109-0030
e-mail: ssgopal@umich.edu

Abstract

We examine image segmentation using a Hopfield neural network. Image segmentation is posed as an optimization problem, and is correlated with the energy function of the neural network. By carefully designing the optimization criterion for segmentation, it is possible to identify the bias inputs and the interconnection weights of the corresponding neural network. We provide a general framework for the design of the optimization criterion, which consists of a component based on the observed image, and another component based on an a priori image model. As an application, we consider a smoothness constraint for the segmented image as our a priori information, and solve a gray-level based segmentation problem. The feasibility of using the neural network architecture based on this optimization criterion for the segmentation of masses in mammograms is demonstrated.

1. Introduction

An important step in computer aided diagnosis in medicine is automated image analysis. Image segmentation is the first step in image analysis. The variety of image segmentation techniques available in the literature encompass those based on simple grey level thresholding, statistical techniques based on probabilistic models for the image formation process, and textured image segmentation via stochastic image models.

Recently several researchers [1, 2, 3, 4] have investigated the use of neural networks for image segmentation. In [1], Amartur and Piraino clustered feature vectors extracted from a magnetic resonance (MR) image using a neural network which minimized the Mahalanobis distance between the feature vectors. Although this approach worked well in the examples shown, it led to sub-optimal image segmentation. This is because pixels in general are spatially correlated and the approach presented in [1] did not incorporate any spatial information. Jamison and Schalkoff [2] incorporated spatial information within a neural network architecture to perform relaxation labeling. Relaxation labeling [5], however, suffers from the drawback that it makes little or no use of any observations (such as the grey level intensities of the pixels) except, perhaps, to initialize an iterative labeling algorithm. Relaxation labeling remains a somewhat *ad hoc* technique that is often used to provide an improvement of labels that have been assigned using some other labeling method.

In this paper we consider optimal image segmentation via pixel classification using a Hopfield neural network. We model the pixel classification problem as an optimization problem and use a neural network for minimizing an energy function which is formulated based on the optimization problem. Our approach differs from those discussed above in that, it not only uses observed data, but also models the local spatial interactions between pixel classes. It is well known that neighboring pixels in an image are "similar" in some sense, a fact exploited in several different statistical image segmentation techniques. Neural network based segmentation using such local spatial interactions, however, has remained a rarity so far.

2. Pixel Classification as an Optimization Problem

Let us consider the pixel classification problem from an optimization standpoint. Frequently the available input is a set of measurements, usually the grey level values of pixels from an $N \times N$ image. Other measurements such as edge information, or local correlation values can be derived from the observed grey scale

0-7803-4122-8/97 $10.00©1997 IEEE

values and be grouped together to constitute a feature vector for each discrete pixel location within the image. Let $\mathbf{y}^i$ denote an ordered vector of M such features associated with the ith pixel in the observed image with $1 \leq i \leq N^2$. We assume the existence of L underlying classes and each pixel is considered to belong to one of these classes. For our purposes it is assumed that the number of classes in the image is known. The pixel classification problem can be stated as follows. Given a known number of classes, assign each pixel in the observed image to one of these L classes.

A commonly used technique that uses feature vectors for segmentation is the K-means algorithm, in which pixels with similar feature vectors are assigned to the same class. The similarity measure is defined as the Euclidean distance to the cluster center vectors, $\mathbf{c}^l$, $l = 1, \ldots, L$, which are updated at each iteration of the algorithm as the average of the feature vectors belonging to class l. A drawback of the K-means algorithm when applied to image segmentation is that it does not have any mechanism to incorporate *a priori* image models.

In this paper, for an L-class segmentation problem, we define an optimization function as

$$\Psi(\mathbf{x}^1, .., \mathbf{x}^{N^2}) = \sum_{i=1}^{N^2} \psi(\mathbf{y}^i, \mathbf{c}^1, .., \mathbf{c}^L, \mathbf{x}^i) - \lambda f(\mathbf{x}^1, .., \mathbf{x}^{N^2}) \ . \quad (1)$$

Here $\mathbf{x}^i$ is an $L \times 1$ label vector such that $x_l^i \in \{0, 1\}$, and $\sum_l x_l^i = 1$ for $1 \leq i \leq N^2, 1 \leq l \leq L$. In Eqn. (1), the function $f(\mathbf{x}^1, .., \mathbf{x}^{N^2})$ models the *a priori* information about the underlying image, with λ representing a scalar constant which denotes the degree to which the *a priori* information is emphasized with respect to the data dependent term $\psi(\mathbf{y}^i, \mathbf{c}^1, .., \mathbf{c}^L, \mathbf{x}^i)$. An optimal solution to the classification problem is a configuration which minimizes $\Psi(\mathbf{x}^1, .., \mathbf{x}^{N^2})$. The cluster centers $\{\mathbf{c}^l, 1 \leq l \leq L\}$ are estimated as a by-product of the solution to the optimization problem. Note that the optimization function specified by Eqn. (1) is very general in the sense that there are no restrictions on either the feature vectors $\mathbf{y}^i$ or the *a priori* label information $f(\mathbf{x}^1, .., \mathbf{x}^{N^2})$.

As an example we defined $\psi(.)$ as

$$\psi(\mathbf{y}^i, \mathbf{c}^1, .., \mathbf{c}^L, \mathbf{x}^i) = \sum_{l=1}^{L} \|(\mathbf{y}^i - \mathbf{c}^l)\|_2^2 x_l^i \quad (2)$$

and the prior term as

$$f(\mathbf{x}^1, .., \mathbf{x}^{N^2}) = \sum_{j \in \eta_i} \mathbf{x}^{j^T} \mathbf{x}^i \ , \quad (3)$$

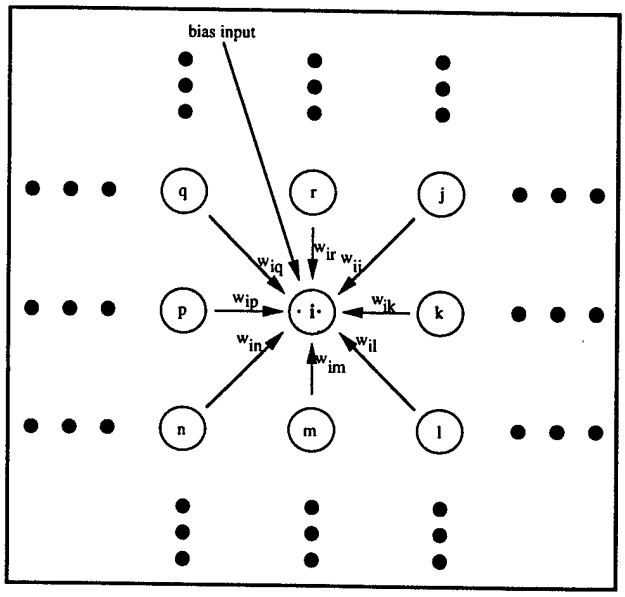


Figure 1. General architecture of a Hopfield neural network.

where η_i is the set of all pixels which are neighbors of pixel i and T denotes the transpose of a vector. For a 1st order neighborhood η_i includes the 4 nearest neighbors of pixel i, for a 2nd order neighborhood η_i includes all the 8 pixels surrounding pixel i, and so on.

3. Network Architecture

Hopfield neural networks [6] have been widely applied to solve an optimization problem such as the one described in the previous section. Figure 1 illustrates the general architecture of a Hopfield network. Each neuron in the Hopfield network represents a pixel location in the image. The generalized energy function associated with this Hopfield network [6] is given by

$$E = -\sum_{i=1}^{N^2} \sum_{j=1}^{N^2} \omega_{i,j} v_i v_j - \sum_{i=1}^{N^2} b_i v_i \ . \quad (4)$$

Here v_i denotes the output of the ith neuron, $\omega_{i,j}$ the strength of the interconnection weight between the ith and jth neuron, and b_i denotes the bias input to the ith neuron. The Hopfield network has the inherent property that as the neuronal outputs evolve over time the network tends to move to a state of lower energy. We can use this inherent property of a Hopfield network for solving the L-class pixel classification problem based on the optimization function defined in Eqn. (1). The interconnection weights and the bias inputs to the neurons are obtained by equating Eqns. (1) and (4) with

$\psi(.)$ and $f(.)$ defined by Eqns. (2) and (3) respectively. For example, let η_i denote the set of pixels within a 3x3 window centered around the ith pixel. Then the *a priori* information is that the label at the ith pixel is more likely to be similar to the labels of pixels within the 3x3 window centered around the ith pixel. The interconnection weights and the bias inputs are given by

$$\omega_{i,j} = 2\lambda \; \forall j \in \eta_i, j \neq i \; , \tag{5}$$

and

$$b_l^i = \frac{1}{L}\sum_{l=1}^{L}\delta_l^i - \frac{8\lambda}{L} - \delta_l^i \tag{6}$$

respectively, where $\delta_l^i \equiv \left\|(\mathbf{y}^i - \mathbf{c}^l)\right\|_2^2$. The input-output relationship for each neuron in the network is defined by

$$v_i = \begin{cases} 1 & \text{if net input to neuron } i \text{ is} > 0 \\ 0 & \text{if net input to neuron } i \text{ is} \leq 0 \end{cases} \tag{7}$$

The overall iterative algorithm can be described as follows:

(1) Initialize cluster centers $\{\mathbf{c}^l\}$ and neuron outputs v_i to some arbitrary values.

(2) At each iteration k and for each neuron i:
(a) compute the net input to the neuron using Eqns. (5) and (6),
(b) compute the output $v_i^{(k)}$ of the neuron using Eqn. (7).

(3) If $v_i^{(k)} \neq v_i^{(k-1)} \; \forall i$ go to step (2). Otherwise go to step (4).

(4) Compute new cluster centers $\{\mathbf{c}^l\}$ using $v_i^{(k)}, i = 1, .., N^2$. Go to step (2).

Steps (2)-(4) are repeated until the cluster centers do not change.

4. Experimental Results and Discussion

We have applied the algorithm described in the previous section to segment masses in mammograms. Regions of interest (ROIs) which included the masses were extracted from the mammograms [7]. In our preliminary study, we used simply the grey levels of the pixels as the input feature. Each pixel of the ROI image was classified as belonging to either the mass or background tissue using the neural network described above. We present two such examples in Figures 2 and 3. Shown are the respective ROI images along with the segmentation results obtained using the neural network with

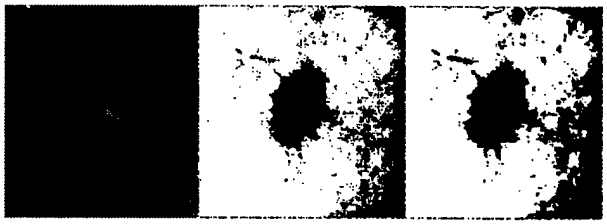


Figure 2. Two-class segmentation results - Example 1: original ROI image (left), and segmentation results using neural network with $\lambda = 0$ (center) and $\lambda = 1$ (right).

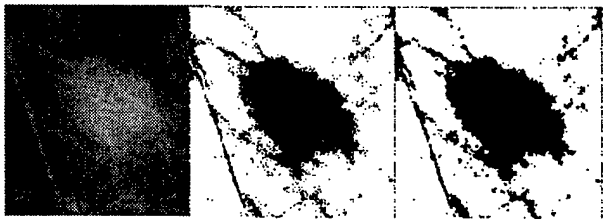


Figure 3. Two-class segmentation results - Example 2: original ROI image (left), and segmentation results using neural network with $\lambda = 0$ (center) and $\lambda = 2$ (right).

$\lambda = 0$ (*i.e.* no *a priori* information), 1 (in Figure 2) and 2 (in Figure 3). As expected, the neural network with λ values of 1 and 2 was observed to yield locally smooth segmentation in contrast to $\lambda = 0$.

In a second study, we extracted a 4-element feature vector for each pixel of the ROI image using a technique described in [8]. Each pixel of the ROI image was first classified into one of three possible classes using the neural network. The region corresponding to the mass was then extracted from the segmented image. Results from this second study are shown in Figures 4-6 for different values of λ. In these figures, the outline or edge of the mass is shown superimposed on the original ROI image. A close observation of Figures 4-6 indicates that the neural network based algorithm incorporating *a priori* information yields improved segmentation and results in better mass boundary delineation. The improved mass boundary definition is expected to provide more accurate estimates of mass characteristics which in turn may lead to more accurate classification of the mass into malignant or benign categories.

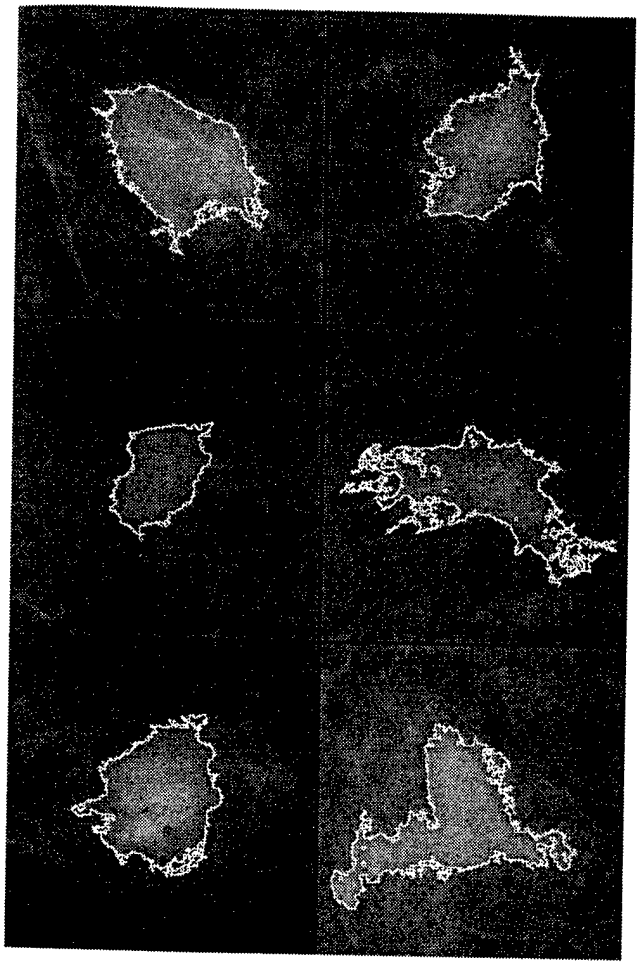


Figure 4. Three-class segmentation results on 6 different ROI images: Outline of the mass segmented using the neural network with $\lambda = 0$ superimposed on the original ROI images.

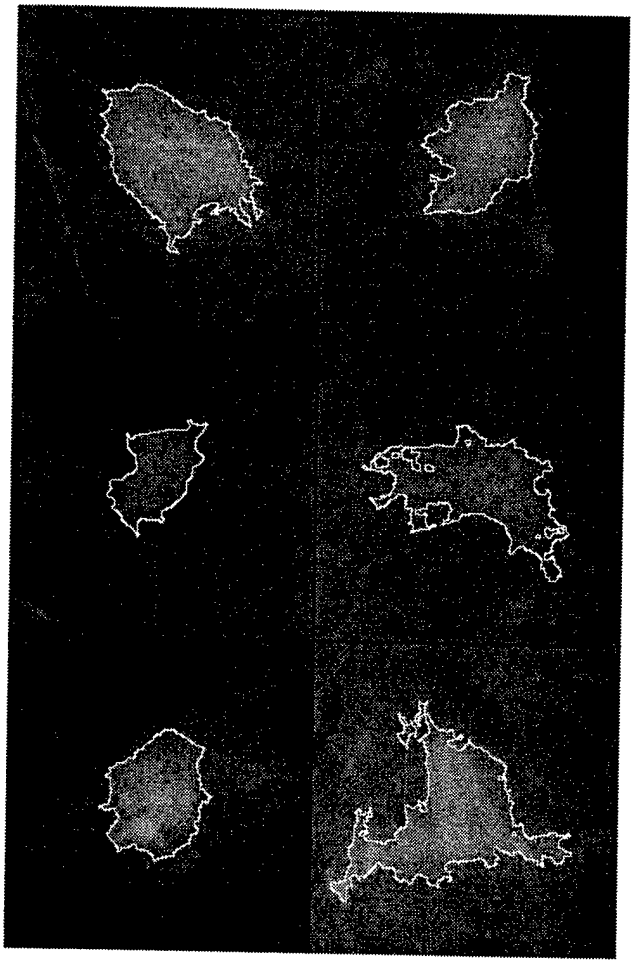


Figure 5. Three-class segmentation results on 6 different ROI images: Outline of the mass segmented using the neural network with $\lambda = 0.1$ superimposed on the original ROI images.

5. Summary and Conclusions

In this study we examined image segmentation via pixel classification using an artificial neural network. Pixel classification was posed as an optimization problem. We provided a general framework for the formulation of an optimization function for pixel classification. By correlating this optimization function with the generalized energy function of the Hopfield network, we defined the bias inputs and the interconnection weights of the network. For simplicity, the optimization criterion we considered here was the Euclidean distance between the feature vectors and the cluster centers. To ensure optimal segmentation we incorporated *a priori* information about the observed image. We demonstrated the feasibility of this approach by a limited study in which masses were successfully segmented from the surrounding mammographic background.

The neural network we have proposed incorporates *a priori* information about the image which results in superior image segmentation. The framework presented in this paper is general enough to handle (i) multidimensional feature vectors, (ii) different forms for the data dependent term of the optimization function described by Eqn. (1). and (iii) any form of user-specified *a priori* information via the prior term in Eqn. (1). At present, our method for choosing an appropriate value for the prior parameter λ is based on visual evaluation of the resulting segmentation. We are investigating statistical and constrained approaches for choosing an

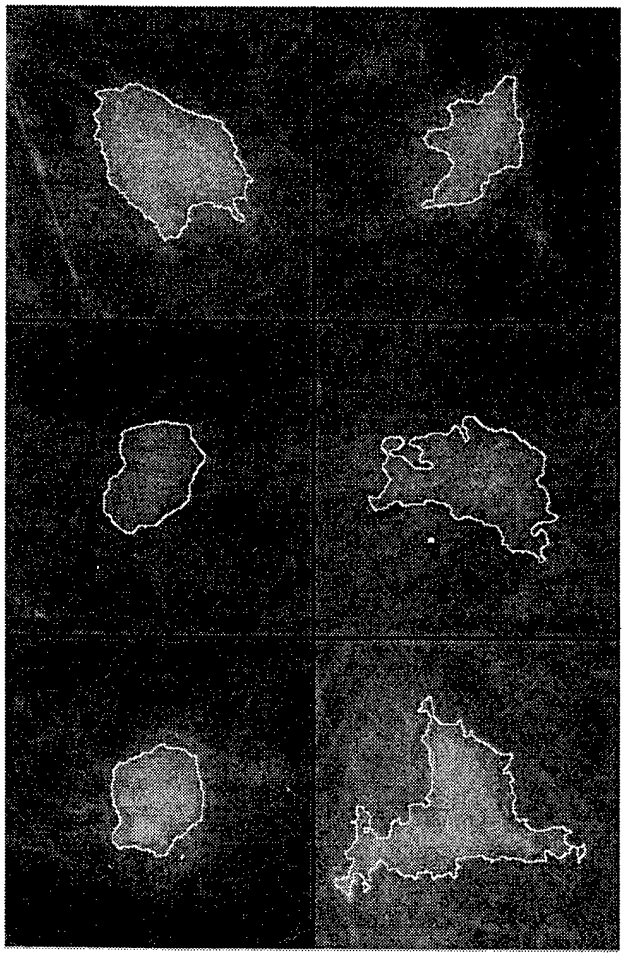


Figure 6. Three-class segmentation results on 6 different ROI images: Outline of the mass segmented using the neural network with $\lambda = 0.25$ superimposed on the original ROI images.

optimal value for this parameter. Directions of future research include studies with a large database, the investigation of appropriate priors for pixel classification, and the formulation of automated methods for choosing appropriate prior parameters.

Acknowledgments

This work is supported by USPHS grant CA 48129, USAMRMC grant DAMD 17-96-1-6254, and a Career Development Award (B. S.) from the USAMRMC (DAMD 17-96-1-6012). The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment or product of any companies mentioned in the publication should be inferred.

References

[1] S. C. Amartur, D. Piraino, and Y. Takefuji, "Optimization neural networks for the segmentation of magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 11, pp. 215-220, 1992.

[2] T. A. Jamison and R. J. Schalkoff, "Image labeling: a neural network approach," *Image and Vision Computing*, vol. 6, pp. 203-213, 1988.

[3] S. W. Lu and H. Xu, "Textured image segmentation using autoregressive models and artificial neural networks," *Pattern Recognition*, vol. 28, pp. 1807-1817, 1995.

[4] S. Haring, M. A. Viergever and J. N. Kok, "Kohonen networks for multiscale image segmentation," *Image and Vision Computing*, vol. 12, pp. 339-344, 1994.

[5] A. Rosenfeld, R.A. Hummel, S.W. Zucker; "Scene labeling by relaxation operations," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-6, no. 6, pp:420-433, 1976.

[6] J. J. Hopfield and D. W. Tank, "Neural computation of decisions in optimization problems," *Biological Cybernetics*, vol. 52, pp. 141-152, 1985.

[7] N. Petrick, H. P. Chan, B. Sahiner, D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *IEEE Transactions on Medical Imaging*, vol. 15, pp. 59-67, 1996.

[8] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M.A. Helvie, D.D. Adler, and M.M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Medical Physics*, vol. 23, pp. 1671-1684, 1996.

Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size

Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Robert F. Wagner*, Lubomir Hadjiiski

Department of Radiology, University of Michigan, Ann Arbor, MI 48109-0904
*Center for Devices and Radiological Health, FDA, Rockville, MD 20857

ABSTRACT

In computer-aided diagnosis (CAD), a frequently-used approach is to first extract several potentially useful features from a data set. Effective features are then selected from this feature space, and a classifier is designed using the selected features. In this study, we investigated the effect of finite sample size on classifier accuracy when classifier design involves feature selection. The feature selection and classifier coefficient estimation stages of classifier design were implemented using stepwise feature selection and Fisher's linear discriminant analysis, respectively. The two classes used in our simulation study were assumed to have multidimensional Gaussian distributions, with a large number of features available for feature selection. We investigated the effect of different covariance matrices and means for the two classes on feature selection performance, and compared two strategies for sample space partitioning for classifier design and testing. Our results indicated that the resubstitution estimate was always optimistically biased, except in cases where too few features were selected by the stepwise procedure. When feature selection was performed using only the design samples, the hold-out estimate was always pessimistically biased. When feature selection was performed using the entire finite sample space, and the data was subsequently partitioned into design and test groups, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, number of available samples, and their statistical distribution. All hold-out estimates exhibited a pessimistic bias when the parameters of the simulation were obtained from texture features extracted from mammograms in a previous study.

Keywords: feature selection, linear discriminant analysis, effects of finite sample size, computer-aided diagnosis

1. INTRODUCTION

A common problem in computer-aided diagnosis (CAD) is the lack of a large number of image samples to design a classifier and to test its performance. The effect of finite sample size on the classification accuracy is therefore an important research topic. In order to treat its specific components, previous studies have mostly ignored the feature selection component of this problem, and assumed that the features used in the classifier were fixed.[1-4] However, in many CAD algorithms, feature selection is a necessary first step. This paper addresses the effect of finite sample size on classification accuracy when the classifier design involves feature selection.

In classifier design, the resubstitution and hold-out estimates are commonly used to assess the accuracy of the classifier. To obtain the resubstitution estimate, the classifier is designed using a number of training samples, and the same samples are then applied to the classifier to yield the distribution of the output decision variable for the training group. The resubstitution performance of the classifier is then measured (e.g., by computing the area under the receiver operating characteristic curve, or by evaluating the probability of misclassification) using this distribution. To obtain the hold-out estimate, the classifier is designed in a similar way, except that an independent set of test samples are applied to the classifier to yield the distribution of the output decision variable for the test group. As the number of training samples increases, both of these estimates approach the true classification accuracy, which is the accuracy of a classifier designed with the full knowledge of the sample distributions. When the training sample size is finite, it is known that, on average, the resubstitution estimate of classifier accuracy is optimistic. In other words, it has a higher expected value than the performance obtained with an infinite design sample set, which is the true classification accuracy. Similarly, on average, the hold-out estimate is pessimistic. When classifier design is limited by the availability of design samples, it is important to obtain a conservative (or pessimistic) performance estimate, which provides a lower bound on the classification accuracy.

In CAD literature, different methods have been used to estimate the classifier accuracy when the classifier design involves feature selection. In a few studies, only the resubstitution estimate was provided.[5] In some studies, the researchers partitioned the samples into training and test groups at the beginning of the study, performed both feature selection and

Part of the SPIE Conference on Image Processing • San Diego, California • February 1999
SPIE Vol. 3661 • 0277-786X/99/$10.00

classifier parameter estimation using the training set, and provided the hold-out performance estimate.[6] Several other studies used a mixture of the two methods: The entire sample space was used as the training set at the feature selection step of classifier design, but once the features were chosen, the hold-out or leave-one-out methods were used to measure the accuracy of the classifier.[7-12] To our knowledge, it has not been reported whether this latter method provides an optimistic or pessimistic estimate of the classifier performance.

This paper describes a simulation study that investigates the effect of finite sample size on classifier accuracy when classifier design involves feature selection. We chose to focus our attention on stepwise feature selection in linear discriminant analysis (stepwise linear discriminant analysis) since this is a simple and common feature selection and classification method. The class distributions were assumed to be multivariate Gaussian. We studied the effect of different covariance matrices and means on feature selection performance. We compared the bias of the classifier when feature selection was performed on the entire sample space, and on the design samples alone. The effects of sample size, number of available features, and parameters of stepwise feature selection on classifier bias were examined.

2. METHODS

To evaluate the effect of sample size on feature selection and classifier bias, we studied the problem of stepwise linear discriminant analysis in two stages. The first stage is stepwise feature selection, and the second stage is the estimation of linear discriminant coefficients for the selected feature subset.

2.1. Stepwise Feature Selection

Stepwise feature selection iteratively enters features into or removes features from the group of selected features based on a feature selection criterion.[13] In our study, we used Wilks' lambda, which is defined as the ratio of within-group sum of squares to the total sum of squares of the discriminant scores, as the feature selection criterion. At the feature entry step of the stepwise algorithm, an F value is computed for each feature based on the ratio of the Wilks' lambda before and after the feature is entered into the pool of already selected features. The feature with the largest F value is entered into the selected feature pool if the F value is larger than a threshold F_{in}. At the feature removal step, the features are tested for removal one at a time from the selected feature pool, the F values are computed, and the feature with the smallest F value is removed from the selected feature pool if the F value is smaller than a threshold F_{out}. The algorithm terminates when no more features can satisfy the criteria for either entry or removal. The number of features selected therefore increases, in general, when F_{in} or F_{out} are reduced.

2.2. Estimation of Linear Discriminant Coefficients

As a by-product of the stepwise feature selection procedure used in our study, the coefficients of a linear classifier that classifies its design samples using the selected features are also computed. However, in this study, the design samples used in the stepwise feature selection step of classifier design may be different from those used in the estimation of classifier coefficients. Therefore, we implemented the stepwise feature selection and the classifier coefficient estimation components of our classification scheme separately.

Let Σ_1 and Σ_2 denote the k-by-k covariance matrices of samples belonging to class 1 and class 2, and let $\mu_1 = (\mu_1(1), \mu_1(2), \ldots, \mu_1(k))$ denote their mean vectors. For an input vector X, the linear discriminant classifier output is defined as

$$h(x) = \frac{1}{2}(\mu_2 - \mu_1)^T \Sigma^{-1} X + \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2), \quad (1)$$

where $\Sigma=(\Sigma_1+\Sigma_2)/2$. The linear discriminant classifier is the optimal classifier when the two classes have a multivariate Gaussian distribution with equal covariance matrices.

For the class separation measures considered in this paper (refer to Section 2.3), the constant term $(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)/2$ in Eq. (1) is irrelevant. Therefore, the classifier design can be viewed as the estimation of k parameters of the vector $(\mu_2 - \mu_1)^T \Sigma^{-1}$ using the design samples.

When a finite number of design samples are available, the means and covariances are estimated as the sample means and the sample covariances from the design samples. The substitution of true means and covariances in Eq. (1) by their estimates causes a bias in the accuracy of the classifier. In particular, if the designed classifier is used for the classification of design samples, then the performance is optimistically biased, and if the classifier is used for classifying test samples that are independent from the design samples, then the performance is pessimistically biased.

2.3. Measures of Class Separation

2.3.1. Infinite sample size

When an infinite sample size is available, the class means and covariance matrices can be estimated without bias (i.e., these quantities can be assumed to be known). In this case, we used the Mahalanobis distance $\Delta(\infty)$, or the area $A_z(\infty)$ under the receiver operating characteristic (ROC) curve as measures of classifier accuracy. The infinity sign in parentheses reflects the fact that the distance is computed using the true means and covariance matrices, or, equivalently, using an infinite number of samples.

Assume that the two classes with a multivariate Gaussian distribution with equal covariance matrices have been classified using Eq. (1). Since Eq. (1) is a linear function of the feature vector X, the classifier outputs for class 1 and class 2 will be Gaussian. Let m_1 and m_2 denote means of the classifier output for the normals and the abnormals, respectively, and let s_1^2 and s_2^2 denote the variances for the two classes. With $\Delta(\infty)$ defined as

$$\Delta(\infty)=(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1), \tag{2}$$

it can easily be shown that

$$m_2-m_1=s_1^2=s_2^2=\Delta(\infty). \tag{3}$$

The quantity $\Delta(\infty)$ is referred to as the Mahalanobis distance between the two classes. It is the Euclidean distance between the two classes, normalized to the common covariance matrix.

In particular, if Σ is an k-by-k diagonal matrix with $\Sigma_{i,i}=\sigma^2(i)$, then

$$\Delta(\infty)=\sum_{i=1}^{k}\delta(i), \tag{4}$$

where

$$\delta(i)=[\mu_2(i)-\mu_1(i)]^2/\sigma^2(i) \tag{5}$$

is the squared signal-to-noise ratio of the difference of the means between the two classes for the i^{th} feature.

Using Eq. (3), and the normality of the classifier outputs, it can be shown that[14]

$$A_z(\infty)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\sqrt{\Delta/2}}e^{-t^2/2}dt \tag{6}$$

2.3.2. Finite sample size

When a finite sample size is available, the means and covariances of the two class distributions were estimated as the sample means and the sample covariances using the training samples, and the classifier outputs for the training and test samples were computed using Eq. (1). The accuracy of the classifier was measured by receiver operating characteristic (ROC) methodology.[15,16] The discriminant scores for samples belonging to class 1 and class 2 were used as decision variables in the LABROC1 program, which provided the ROC curve based on maximum likelihood estimation.

2.4. Simulation conditions

For our simulations, we assumed that the two classes have a multivariate Gaussian distribution with equal covariance matrices, and different means. The number of available features was M=100. We generated a sample size of N_s samples from each class using a random number generator. The sample space was randomly partitioned into N_t training samples and N_s-N_t test samples per class. For a given sample space, we used several different values for N_t in order to study the effect of the design sample size on classification accuracy. In order to reduce the variance of the classification accuracy

estimate, a given sample space was independently partitioned 20 times into N_t training samples and N_s-N_t test samples per class, and the classification accuracy using these 20 partitions was averaged. The procedure described above was referred to as an experiment. For each simulation condition described below, 50 statistically independent experiments were performed, and the results were averaged.

Two methods for feature selection were considered. In the first method, the entire sample space was used for feature selection. In other words, the entire sample space was treated as a training set at the feature selection step of classifier design. Before the coefficient estimation step of classifier design, the sample space was partitioned into training and test groups. The training group was used for classifier coefficient estimation, and the resubstitution and hold-out performances were estimated by applying the training and test groups to the designed classifier, respectively. In the second method, sample set partitioning was performed before feature selection. In other words, both feature selection and coefficient estimation were performed only on the training set.

Case 1: Comparison of correlated and diagonal covariance matrices

Case 1.a

In this simulation condition, the 100X100 covariance matrix Σ was chosen to have a block-diagonal structure

$$\Sigma = \begin{bmatrix} A & 0 & 0 & \cdots & 0 \\ 0 & A & 0 & \cdots & 0 \\ 0 & 0 & A & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & A \end{bmatrix}$$

where the 10X10 matrix A was defined as

$$A = \begin{bmatrix} 1 & 0.8 & 0.8 & \cdots & 0.8 \\ 0.8 & 1 & 0.6 & \cdots & 0.6 \\ 0.8 & 0.6 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0.6 \\ 0.8 & 0.6 & \cdots & 0.6 & 1 \end{bmatrix}$$

and $\Delta\mu(i)$=0.1732 for all i. Using (2), the Mahalanobis distance is computed as $\Delta(\infty)$=3.0, and $A_z(\infty)$=0.89.

Case 1.b

The features in Case 1.a can be transformed into a set of uncorrelated features using a linear transformation, which is called the orthogonalization transformation. The linear orthogonalization transformation is defined by the eigenvector matrix of Σ, so that the covariance matrix after orthogonalization is diagonal. After the transformation, the new covariance matrix turns out to be the identity matrix, and the new mean vector is

$$\Delta\mu(i) = \begin{cases} 0.5477 & \text{if } i \text{ is a multiple of 10} \\ 0 & \text{otherwise} \end{cases}$$

Since a linear transformation will not affect the separability of the two classes, the Mahalanobis distance is the same as in Case 1.a, i.e., $\Delta(\infty)$=3.0.

Case 2: Simulation of a possible condition in CAD

In order to simulate covariance matrices and mean vectors that one may encounter in CAD, we used texture features extracted from patient mammograms in a previous study, which aimed at classifying regions of interest (ROIs) on mammograms as malignant or benign.[7] Ten different spatial gray level dependence (SGLD) texture measures were extracted from each ROI at five different distances and two directions. The number of available features was therefore M=100. The transformations that were applied to the ROI before feature extraction, and the formal definition of SGLD features can be found in the literature.[7,17] The means and covariances for each class were estimated from a database of 249 mammograms.

Case 2.a

In this simulation condition, the two classes were assumed to have a multivariate Gaussian distribution with $\Sigma=(\Sigma_1+\Sigma_2)/2$, where Σ_1 and Σ_2 were estimated from the feature samples for the malignant and benign classes. Since the features have different scales, their variances can vary by as much as a factor of 10^6. Therefore, it is difficult to provide an idea about how the covariance matrix is distributed without listing all the entries of the 100X100 matrix Σ. The correlation matrix, which is normalized so that all diagonal entries are unity, is better suited for this purpose. The absolute value of the correlation matrix is shown as an image in Fig. 1. In this image, small elements of the correlation matrix are displayed as darker pixels, and the diagonal elements, which are unity, are displayed as brighter pixels. From Fig. 2, it is observed that some of the features are highly correlated or anticorrelated. The Mahalanobis distance was computed as $\Delta(\infty)=2.4$, which implied $A_z(\infty)=0.86$.

Case 2.b

To determine the performance of a feature space with equivalent discrimination potential, but independent features, we performed an orthogonalization transformation on the SGLD feature space, as explained previously (Case 1.b).

3. RESULTS

Case 1:

Feature selection from the entire sample space

Figs. 2.a and 2.b plot the area A_z under the ROC curve for the resubstitution and hold-out performance estimates versus the inverse of the number of training samples per class, $1/N_t$, for Case 1.a, and Case 1.b, respectively (number of samples per class N_s=100). The F_{in} value was varied between 0.5 and 1.5, and F_{out} was defined as $F_{out}=max[(F_{in}-1),0]$. Fig. 3 is equivalent to Fig. 2.a, except the number of samples per class was increased from N_s=100 to N_s=500 in this figure.

Case 2:

Feature selection from the entire sample space

The area A_z under the ROC curve for the resubstitution and hold-out performance estimates are plotted versus $1/N_t$ in Figs. 4.a and 4.b for Case 2.a, and Case 2.b, respectively (N_s=100). The F_{in} value was varied between 0.5 and 3.0, and F_{out} was defined as $F_{out}=max[(F_{in}-1),0]$. Fig. 5 is equivalent to Fig. 4.a, except the number of samples per class was increased from N_s=100 to N_s=500 in this figure.

Feature selection from training samples alone

Case 2.a was used as an example. The area A_z under the ROC curves versus $1/N_t$ are plotted for N_s=100 and N_s=500 in Figs. 6 and 7, respectively.

4. DISCUSSION

Fig. 2.b demonstrates the potential disadvantage of performing feature selection using the entire sample space. The best possible test performance with infinite sample size for Case 1 is $A_z(\infty)=0.89$. However, in Fig. 2.b, we observe that some of the "hold-out" estimates were as high as 0.92. These estimates were higher than $A_z(\infty)$ because the hold-out samples were excluded from classifier design only in the parameter estimation stage of the design, and were used as training samples in feature selection. When feature selection is performed using a small sample size, some features that are useless for the general population may appear to be useful for the classification of the small number of samples at hand. This was previously demonstrated in the literature by comparing the probability of misclassification based on either a finite sample set or the entire population subject to the constraint that a given number of features were used for classification.[18] In our study, given a small data set, the variance in Wilks' lambda estimates causes some feature combinations to appear more powerful than they actually are. If the data set is partitioned into training and test groups after feature selection, these feature combinations may provide optimistic hold-out estimates.

The observation made in the previous paragraph about feature selection using the entire sample space is not a general rule, however. Figs. 2.a and 4.a show that one does not always run the risk of obtaining an optimistic bias in the hold-out estimate when the feature selection is performed using the entire sample space. For Case 1, the best possible test performance with an infinite sample size is $A_z(\infty)=0.89$, but the best hold-out estimate in Fig. 2.a is $A_z=0.82$. Similarly, for Case 2, the best possible test performance with infinite sample size is $A_z(\infty)=0.86$, but the best hold-out estimate in Fig. 4.a is $A_z=0.84$. The features in both Case 1.a and Case 2.a were correlated. Case 1.b and Case 2.b were obtained from Case 1.a and Case 2.a by applying a linear orthogonalization transformation to the features so that they become uncorrelated. Figs. 2.b and 4.b show that after this transformation is applied, the hold-out estimates can be optimistically biased for small sample size

(N_s=100). This shows that performing a linear combination of features before stepwise feature selection can have a dramatic influence on its performance. This result is somewhat surprising, because the stepwise procedure is known to select a set of features whose linear combination can effectively separate the classes. However, the orthogonalization transformation in this study is assumed to be known *a priori* (i.e., it is not deduced from the available finite sample size), and is applied to the entire feature space of M features, whereas the stepwise procedure only produces combinations of a subset of these features.

Figs. 6 and 7 demonstrate that when feature selection is performed using the training set alone, the hold-out performance estimate is pessimistically biased. This bias decreases as the number of training samples, N_t, is increased.

When F_{in} and F_{out} values were low, the resubstitution performance estimates were optimistically biased for all the cases studied. Low F_{in} and F_{out} values imply that many features are selected using the stepwise procedure. From previous studies, it is known that a larger number of features in classification leads to larger resubstitution bias.[3] On the other hand, when F_{in} and F_{out} values were very high, the number of selected features could be so low that the resubstitution estimate would be pessimistically biased, as can be observed from Fig. 3 (F_{in}=1.5) and Fig. 4.a (F_{in}=3.0). In all of our simulations, for a given number of training samples N_t, the resubstitution estimate increased monotonically as the number of selected features were increased by decreasing F_{in} and F_{out}.

In contrast to the resubstitution estimate, the hold-out estimate for a given number of training samples did not change monotonically as F_{in} and F_{out} were decreased. This can be observed from Fig. 2.a, where the hold-out estimate for F_{in}=1.5 is larger than all other hold-out estimates with different F_{in} values for N_t=25 ($1/N_t$=0.04). However, for N_t=90 ($1/N_t$=0.011), the hold-out estimate for the same F_{in} value is no longer the largest. In Fig. 2.a, the feature selection was performed using the entire sample space. A similar phenomenon can be observed in Fig. 7, where the feature selection is performed using the training samples alone. This means that for a given number of design samples, there is an optimum value for F_{in} and F_{out} (or the number of selected features) that provides the highest hold-out estimate. This is the well-known peaking phenomenon described in the literature,[19] which can be explained as follows. For a given number of training samples, increasing the number of features in the classification has two opposing effects on the hold-out performance. On the one hand, the new features may provide some new information about the two classes, which tends to increase the hold-out performance. On the other hand, the same features increase the complexity of the classifier, which tends to decrease the hold-out performance. Depending on the balance between how much new information the new features provide and how much the complexity increases, the hold-out performance may increase or decrease when the number of features is increased.

In this study, the number of available features was fixed at M=100. The number of samples per class was N_s=100 in most of the simulations. However, in three of our simulation conditions, we used N_s=500, which meant that the total number of samples was ten times that of available features. The results of these simulations are shown in Fig. 3 for Case 1, and Figs. 5 and 7 for Case 2. Our first observation concerning these figures is that no hold-out estimates in any of these figures are higher than their respective $A_z(\infty)$ values. This suggests that optimistic hold-out estimates may be avoided by increasing the number of available samples, or, possibly, by decreasing the number of features used for feature selection. A second observation is that, compared to other figures in this study, the relationship between the A_z values and $1/N_t$ is closer to a linear relation. This suggests that it may be possible to obtain $A_z(\infty)$ by fitting a line to the A_z vs. $1/N_t$ curves using linear regression, and finding the y-axis intercept. This is similar to the modified Fukunaga and Hayes technique that we discussed previously in the studies of finite sample size effect on classifier bias.

This study examined only the bias of the mean performance estimates, which were obtained by averaging the estimates from fifty experiments as described in Section 2.4. Another important issue in classifier design is the variance of the individual estimates. The variance provides an estimate of the generalizability of the classifier performance to other design and test samples. We previously studied the variance of performance estimates when the classifier design included the estimation of classifier coefficients, but excluded feature selection.[4,20] The extension of our previous studies to include feature selection is an important further research topic.

5. CONCLUSION

In this study, we investigated the finite-sample performance of a linear classifier that included stepwise feature selection as a design step. We compared the resubstitution and hold-out estimates to the true classification accuracy, which is the accuracy of a classifier designed with the full knowledge of the sample distributions. We compared the effect of partitioning the data set into training and test groups before performing feature selection, and after performing feature

selection. When data partitioning was performed before feature selection, the hold-out estimate was always pessimistically biased. When partitioning was performed after feature selection, i.e., the entire sample space was used for feature selection, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, number of available samples, and their statistical distribution. All hold-out estimates exhibited a pessimistic bias when the parameters of the simulation were obtained from correlated texture features extracted from mammograms in our previous study. The understanding of the performance of the classifier designed with different schemes will allow us to utilize a limited sample set efficiently and to avoid an overly optimistic assessment of the classifier

6. ACKNOWLEDGMENTS

This work is supported by a USPHS Grant No. CA 48129 and a USAMRMC grant (DAMD 17-96-1-6254). B. Sahiner and L. Hadjiiski are also supported by Career Development Awards from the USAMRMC (DAMD 17-96-1-6012 and DAMD 17-98-1-8211). N. Petrick is also supported by a grant from the Whitaker Foundation. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC1 programs.

REFERENCES

1. H.-P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: Quadratic and neural network classifiers," Proc. SPIE Conf. Medical Imaging 3034, 1102-1113 (1997).

2. R. F. Wagner, H.-P. Chan, J. Mossoba, B. Sahiner, and N. Petrick, "Finite-sample effects and resampling plans: Application to linear classifiers in computer-aided diagnosis," Proc. SPIE Conf. Medical Imaging 3034, 467-477 (1997).

3. H.-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Effects of sample size on classifier design for computer-aided diagnosis," Proc. SPIE Conf. Medical Imaging 3338, 845-858 (1998).

4. R. F. Wagner, H.-P. Chan, J. Mossoba, B. Sahiner, and N. Petrick, "Components of variance in ROC analysis of CAD_X classifier performance," Proc. SPIE Conf. Medical Imaging 3338, 859-875 (1998).

5. C.-M. Wu, Y.-C. Chen, and K.-S. Hsieh, "Texture feature for classification of ultrasonic liver images," IEEE Transactions on Medical Imaging 11, 141-152 (1992).

6. P. A. Freeborough and N. C. Fox, "MR image texture analysis applied to the diagnosis and tracking of Alzheimer's disease," IEEE Trans. Medical Imaging 17, 475-479 (1998).

7. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. 25, 516-526 (1998).

8. B. S. Garra, B. H. Krasner, S. C. Horri, S. Ascher, S. K. Mun, and R. K. Zeman, "Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis," Ultrasonic Imaging 15, 267-285 (1993).

9. K. G. A. Gilhuijs and M. L. Giger, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," Medical Physics 25, 1647-1654 (1998).

10. M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," IEEE Trans. Medical Imaging 14, 537-547 (1995).

11. Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer," Radiology 187, 81-87 (1993).

12. V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Medical Physics 19, 1475-1481 (1992).

13. N. R. Draper, *Applied regression analysis*, (Wiley, New York, 1998).

14. A. J. Simpson and M. J. Fitter, "What is the best index of detectability," Psychological Bulletin 80, (1973).

15. C. E. Metz, "ROC methodology in radiologic imaging," Invest Radiol 21, 720-733 (1986).

16. C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Statistics in Medicine 17, 1033-1053 (1998).

17. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Systems Man Cybernetics SMC-3, 610-621 (1973).

18. S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 252-264 (1991).

19. G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Information Theory 14, 55-63 (1968).

20. R. F. Wagner, H.-P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Components of variance in ROC analysis of CAD_X classifier performance: Applications of the bootstrap," Proc. SPIE Conf. Medical Imaging 3661, (in print) (1999).

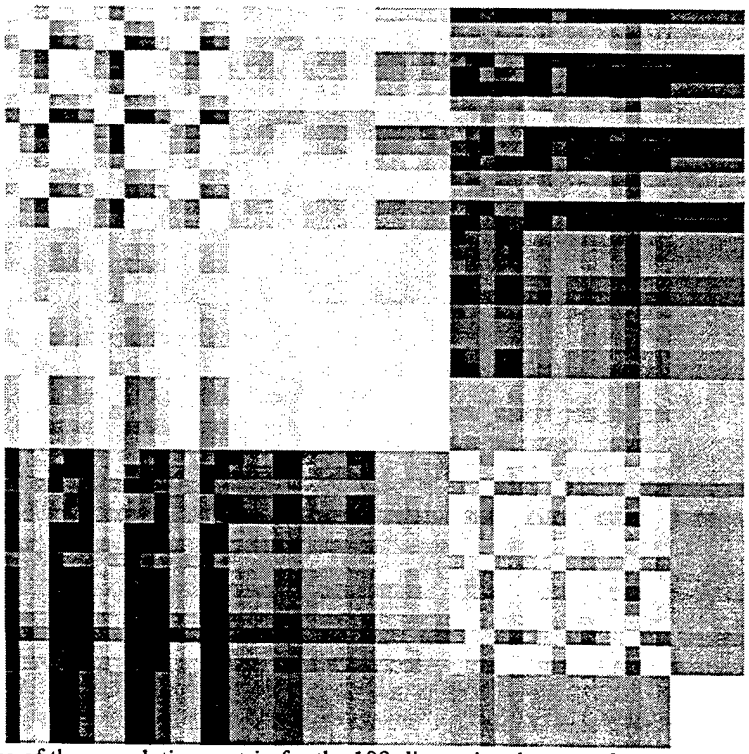


Fig. 1 The absolute value of the correlation matrix for the 100-dimensional texture feature space extracted from 249 mammograms. The covariance matrix corresponding to these features was used in simulation Case 2.a.

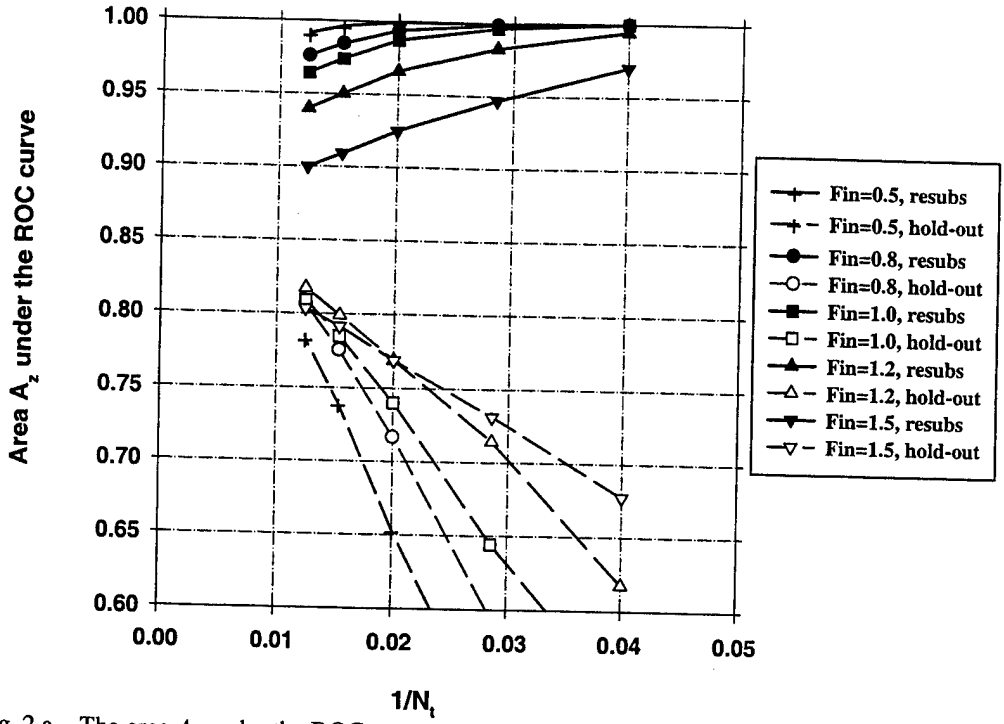


Fig. 2.a The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.a, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

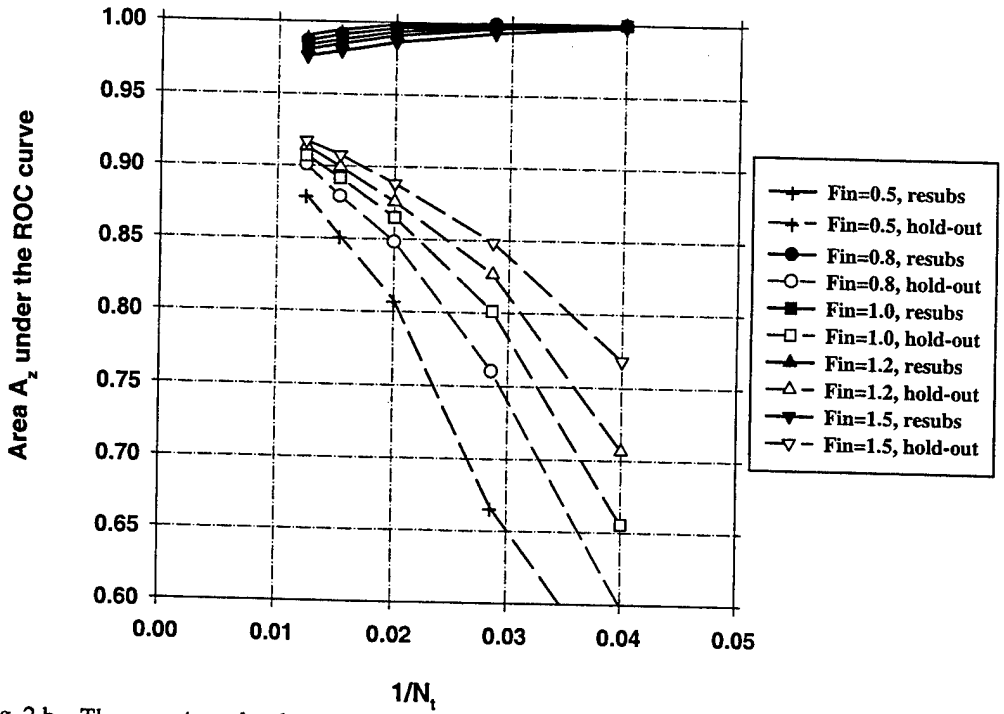


Fig. 2.b The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.b, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

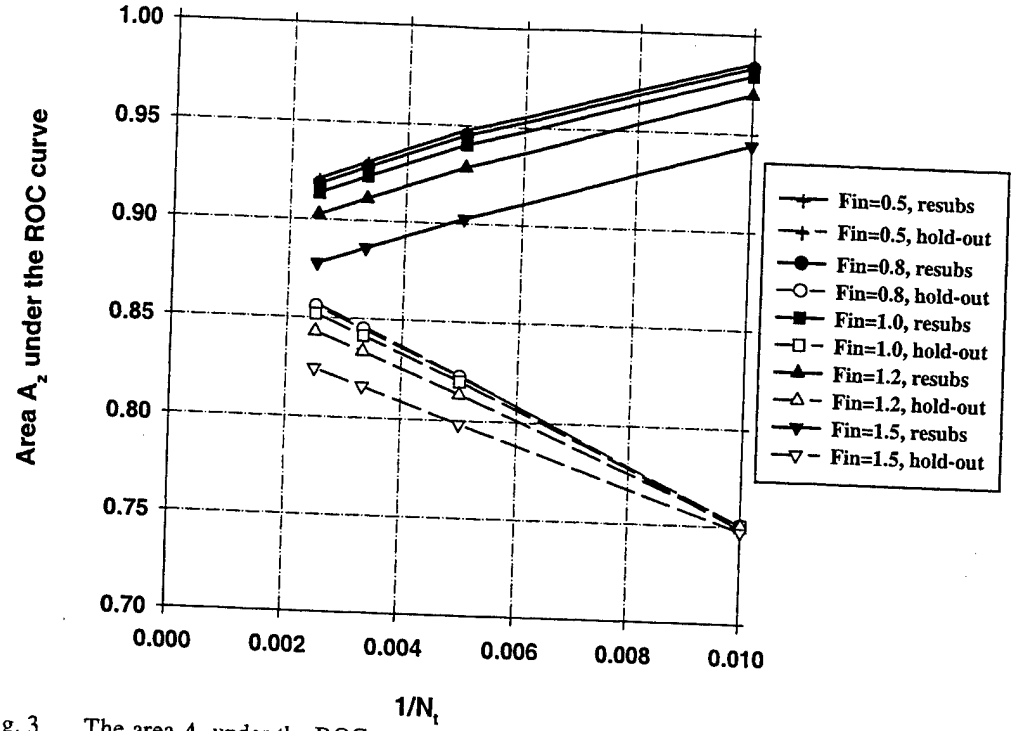


Fig. 3 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.a, feature selection from the entire sample space of 500 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

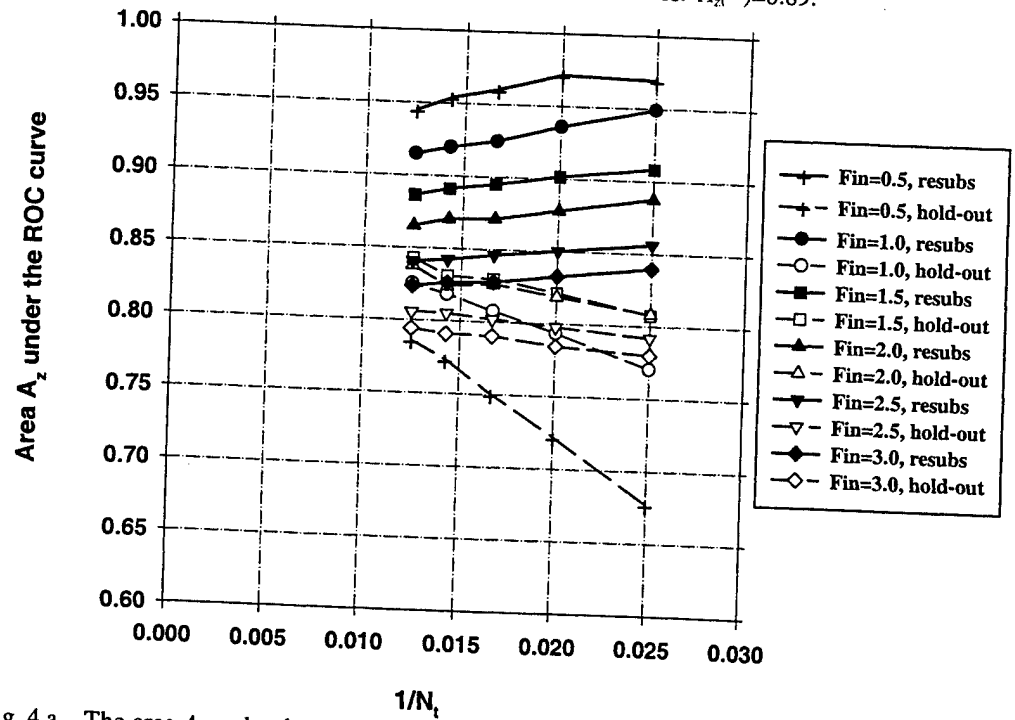


Fig. 4.a The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

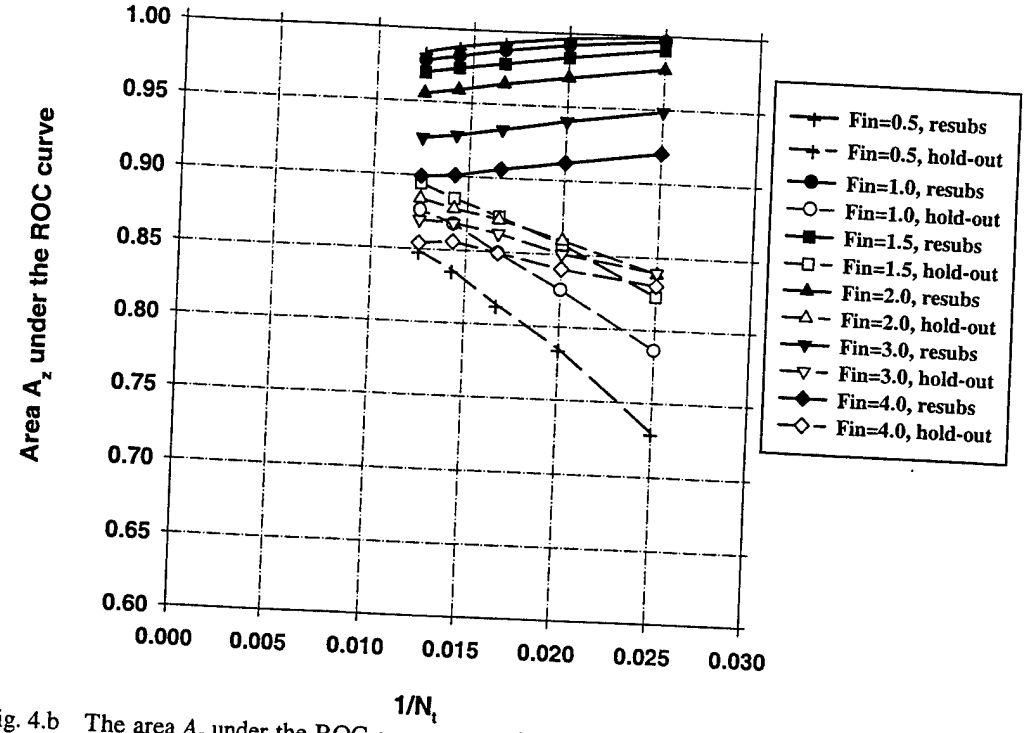


Fig. 4.b The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.b, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

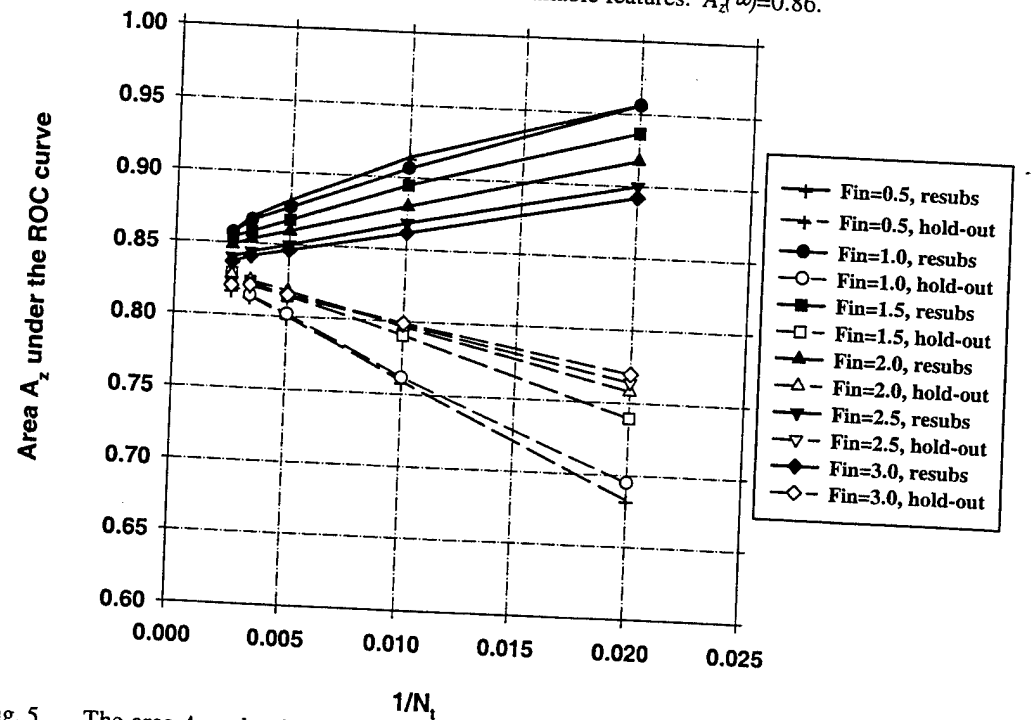


Fig. 5 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from the entire sample space of 500 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

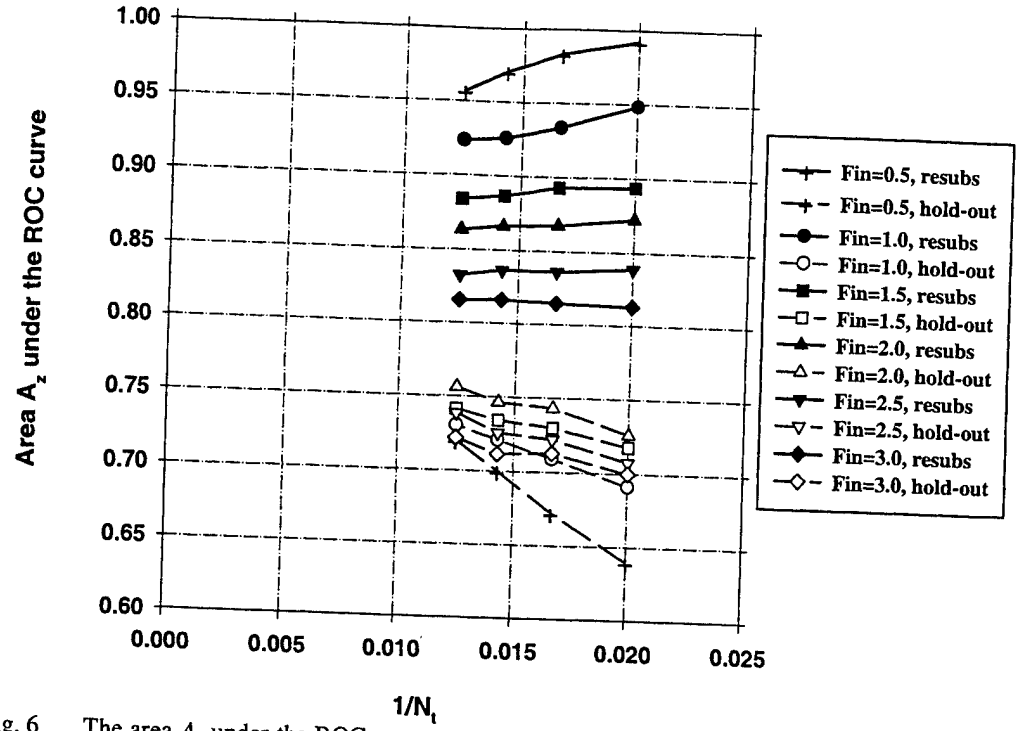


Fig. 6 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from design samples alone (N_s=100). Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

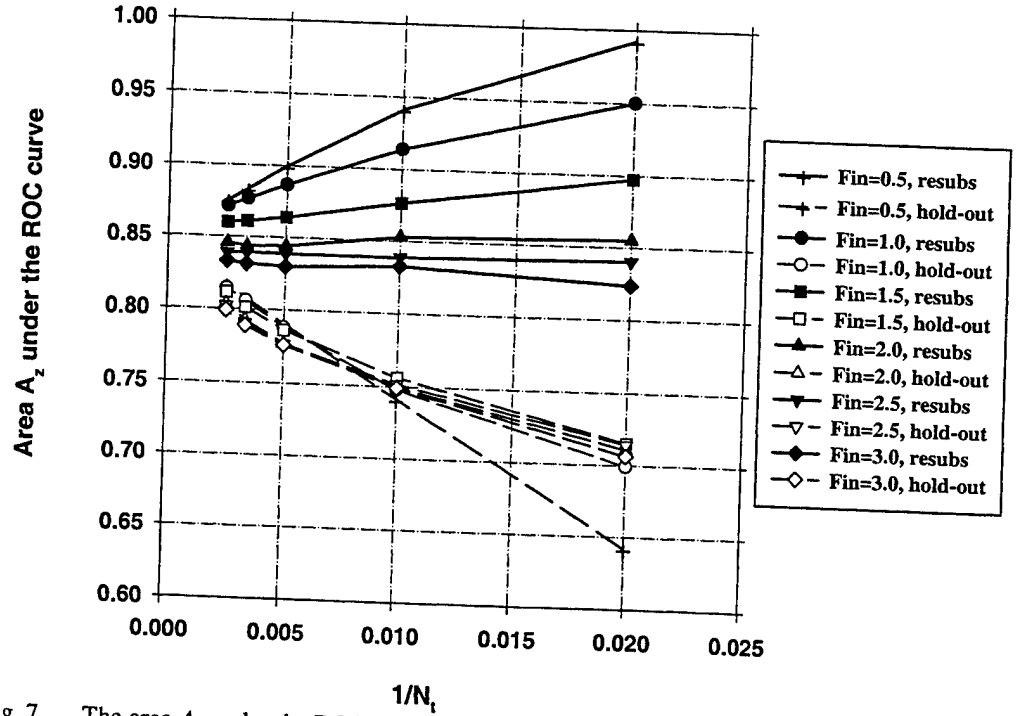


Fig. 7 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from design samples alone (N_s=500). Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms

Lubomir Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Mark Helvie

Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0904

ABSTRACT

A hybrid classifier which combines an unsupervised adaptive resonance network (ART2) and a supervised linear discriminant classifier (LDA) was developed for analysis of mammographic masses. Initially the ART2 network separates the masses into different classes based on the similarity of the input feature vectors. The resulting classes are subsequently divided into two groups: (i) classes containing only malignant masses and (ii) classes containing both malignant and benign or only benign masses. All masses belonging to the second group are used to formulate a single LDA model to classify them as malignant and benign. In this approach, the ART2 network identifies the highly suspicious malignant cases and removes them from the training set, thereby facilitating the formulation of the LDA model. In order to examine the utility of this approach, a data set of 348 regions of interest (ROIs) containing biopsy-proven masses (169 benign and 179 malignant) were used. Ten different partitions of training and test groups were randomly generated using 73% of ROIs for training and 27% for testing. Classifier design including feature selection and weight optimization was performed with the training group. The test group was kept independent of the training group. The performance of the hybrid classifier was compared to that of an LDA classifier alone. Receiver Operating Characteristics (ROC) analysis was used to evaluate the accuracy of the classifier. The average area under the ROC curve (A_z) for the hybrid classifier was 0.81 as compared to 0.78 for LDA. The A_z values for the partial areas above a true positive fraction of 0.9 were 0.34 and 0.27 for the hybrid and the LDA classifier, respectively. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.

1. INTRODUCTION

Mammography is the most effective method for detection of early breast cancer[1]. However, the specificity for classification of malignant and benign lesions from mammographic images is relatively low. Clinical studies have shown that the positive predictive value (i.e., ratio of the number of breast cancers found to the total number of biopsies) is only 15% to 30% [2-3]. It is important to increase the positive predictive value without reducing the sensitivity of breast cancer detection. Computer-aided diagnosis (CAD) has the potential to increase the diagnostic accuracy by reducing the false-negative rate while increasing the positive predictive values of mammographic abnormalities.

Classifier design is an important step in the development of a CAD system. A classifier has to be able to merge the available input feature information and make a correct evaluation. Commonly used classifiers for CAD include linear discriminants (LDA)[4] and backpropagation neural networks (BPN)[5] which have been shown to perform well in lesion classification problems[6-9]. These classifiers are generally designed by supervised training. However, these types of classifiers have limitations dealing with the nonlinearities in the data (in case of LDA) and in generalizability when a limited number of training samples are available (especially BPN). Another classification approach is based on unsupervised classifiers, which cluster the data into different classes based on the similarities in the properties of the input feature vectors. Therefore, unsupervised classifiers can be used to analyze the similarities within the data. However, it is difficult to use them as a discriminatory classifier[16,17].

We propose here a hybrid unsupervised/supervised structure to improve classification performance. The design of this structure was inspired by neural information processing principles such as self-organization, decentralization and generalization. It combines the Adaptive Resonance Theory network (ART2)[14,15] and the LDA classifier as a cascade system (ART2LDA). The self-organizing unsupervised ART2 network automatically decomposes the input samples into classes with different properties. The ART2 network performs better compared to conventional clustering techniques in terms of learning speed and discriminatory resolution for the detection of rare events[16,17]. The supervised LDA then classifies the

Part of the SPIE Conference on Image Processing • San Diego, California • February 1999
SPIE Vol. 3661 • 0277-786X/99/$10.00

samples belonging to a subset of classes that have greater similarities. By improving the homogeneity of the samples, the classifier designed for the subset of classes may be more robust.

The ART2LDA design implements both structural and data decomposition. Decomposition is a powerful approach that can reduce the complexity of a problem. Both structural decomposition and data decomposition can improve classification accuracy[10] as well as model accuracy[11]. However, decomposition can also reduce the prediction accuracy due to overfitting the training data. We will demonstrate in this paper that the proposed hybrid structure can deal with the overfitting problem and improve the prediction capabilities of the system.

2. ART2 UNSUPERVISED NEURAL NETWORK

The ART2 is a self organizing system that can simulate human pattern recognition. ART2 was first described by Grossberg[12,13] and a series of further improvements were carried out by Carpenter, Grossberg and co-workers[14,15]. The ART2 network clusters the data into different classes based on the properties of the input feature vectors. The members within a class have similar properties. The process of ART2 network learning is a balance between the plasticity and stability dilemma. Plasticity is the ability of the system to discover and remember important new feature patterns. Stability is the ability of the system to remain unchanged when already known feature patterns with noise are input to the system. The balance between plasticity and stability for the ART2 training algorithm allows fast learning, i.e., rare events can be memorized with a small number of training iterations without forgetting previous events. The more conventional training algorithms such as backpropagation[5] perform slow learning, i.e., they tend to average over occurrences of similar events and require a lot of training iterations.

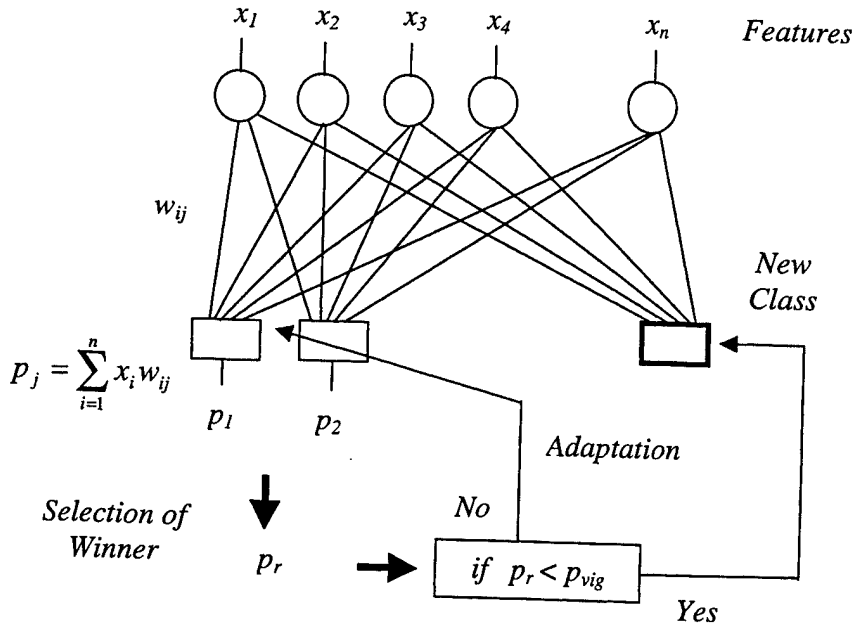


Figure 1. Structure of the ART2 network.

The structure of the ART2 system is shown in Figure 1. It consists of two parts: the ART2 network and the learning stage. Suppose that there are n input features x_i (i=1, ... n) and k classes in the ART2 network. When a new vector is presented to the input of the ART2 network, an activation value p_j for class j is calculated as:

$$p_j = \sum_{i=1}^{n} x_i w_{ij}, \quad j = 1, \ldots, k, \tag{1}$$

where w_{ij} is the connection weight between input i and class j. The activation value is a measure of the membership of the particular input feature vector to class j. The higher the value p_j is, the better the input vector matches class j. The maximum value p_r is selected from all p_j ($j = 1, \ldots, k$) to find the best class match.

Furthermore, in order to balance the contribution to the activation value from all feature components, the input feature values applied to the ART2 system are scaled between zero and one[17]. This normalization will allow detection of similar feature patterns even when the magnitudes of the input feature components are very different.

The learning stage of the ART2 system can influence the weights of the selected class or the complete ART2 network structure by adding a new class. An additional parameter, the vigilance, is used to determine the type of learning[14]. The vigilance parameter p_{vig} is a threshold value that is compared to the maximum activation value p_r. If p_r is larger than p_{vig} then the input vector is considered to belong to class r. The adaptation of the weights connected with class r is performed as follows:

$$w_{ir}^{new} = w_{ir}^{old} + \eta (x_i - w_{ir}^{old}) \quad \text{for } i = 1, \ldots, n, \tag{2}$$

where η is a learning rate. The adaptation of the class r weights (Eq. 2), aims at maximization of the p_r value for the particular input vector. In an iterative manner the weights are adjusted so that the produced activation values for similar input vectors will be maximum only for the class to which they belong and these maximum activation values will be higher than p_{vig}.

If the maximum activation value p_r is smaller than p_{vig}, it is an indication that a novelty has appeared and a new class will be added to the ART2 structure. The new weights connecting the input with the new class $(k+1)$ are initialized with the scaled input feature values of this novelty. In this way the activation value p_{k+1} will be maximum ($p_r = p_{k+1}$) and will be higher than p_{vig}, when it is computed for this novelty in further training iterations. The value of the vigilance parameter p_{vig} determines the resolution of ART2. It can be chosen in the range between 0 and 1. If p_{vig} is relatively small, only very different input feature vectors will be distinguished and separated in different classes. If p_{vig} is relatively large the input feature vectors that are more similar will be separated into different classes. The choice of p_{vig} is depends on the particular application.

3. ART2LDA CLASSIFIER

Despite the good performance of ART2 for efficient clustering and detection of novelties, the fast learning approach can cause problems associated with the generalization capability of the system and the correct classification of unknown cases. Supervised classifiers such as linear discriminants or backpropagation neural network classifiers can have better generalization capability than ART2, because they are trained by averaging over similar event occurrences. However, these classifiers do not have the ability to correctly classify rare events.

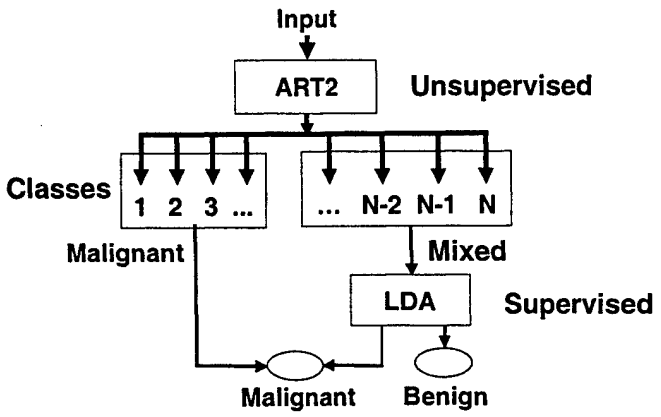


Figure 2. Structure of the ART2LDA classifier.

In order to improve the accuracy and generalization of a classifier, we propose to design a hybrid classifier that combines the unsupervised ART2 network and a supervised LDA classifier. This hybrid classifier (ART2LDA) utilizes the good resolution capability of ART2 and the good generalization capability of LDA. The ART2 network first analyzes the similarity of the sample population and identifies a subpopulation that may be separated from the main population. This will improve the performance of the second-stage LDA if the subpopulation causes the sample population to deviate from a multivariate normal distribution for which LDA is an optimal classifier. Therefore, the ART2 serves as a screening tool to improve the normality of the sample distribution by classifying outlying samples into separate classes.

The structure of the hybrid ART2LDA classifier is shown in Fig. 2. The classes identified by ART2 are labeled to be one of the two types: malignant class or mixed class. A particular class is defined as malignant if it contains only malignant members. It is defined as mixed if it contains both malignant and benign members. The type of a given class is determined based on ART2 classification of the training data set. The ART2 classifies an input sample into either a malignant or a mixed class. Depending on the class type it is determined whether the LDA classifier will be used. If an input sample is classified into a mixed class, the final classification will be obtained based on the LDA classifier, which has been trained by the mixed classes in the training set. However, if an input sample is classified by ART2 into a malignant class then the mass will be considered malignant, without using the LDA classifier. Therefore, in the ART2LDA structure, the ART2 is used both as a classifier and a supervisor.

4. MATERIALS AND METHODS

4.1. Data set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsy at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. Approximately equal number of malignant and benign masses were included. The data set contained 348 mammograms with a mixture of benign (n=169) and malignant (n=179) masses. The visibility of the masses was rated by a radiologist experienced in breast imaging on a scale of 1 to 10, where the rating of 1 corresponds to the most visible

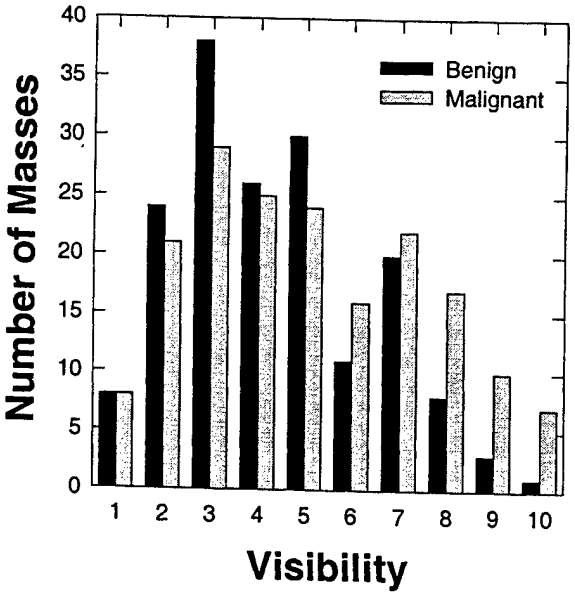


Figure 3. The distribution of the visibility ranking of the masses in the dataset. The ranking was performed by an experienced radiologist. (1: very obvious, 10: very subtle).

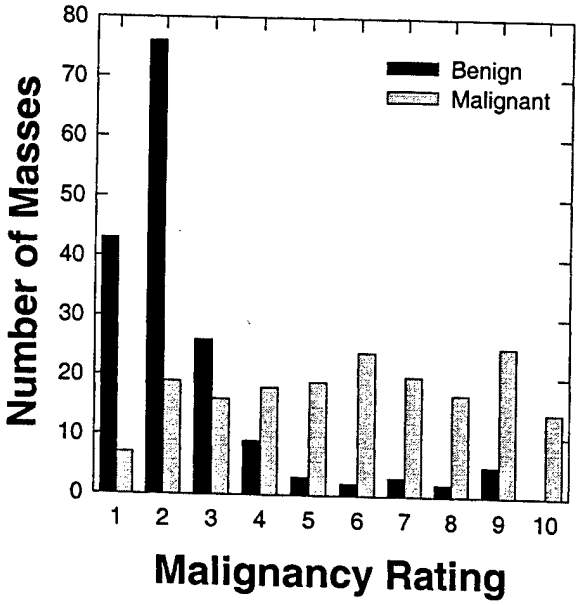


Figure 4. The distribution of the malignancy ranking of the masses in the dataset. The ranking was performed by an experienced radiologist. (1: very likely benign, 10: very likely malignant).

category. The distributions of the visibility rating for both the malignant and benign masses are shown in Fig. 3. The visibility ranged from subtle to obvious for both types of masses. It can be observed that the benign masses tend to be more obvious than the malignant ones. Additionally the likelihood of malignancy for each mass was estimated based on its mammographic appearance. The radiologist rated the likelihood of malignancy on a scale of 1 to 10, where 1 indicated a mass with the most benign appearance. The distribution of the malignancy rating of the masses is shown in Fig. 4.

Three hundred and five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm × 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of –0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0 to 3.5. The remaining 43 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 50 μm × 50 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the OD within the range of 0 to 4 OD units, with a slope of –0.001 OD/pixel value. In order to process the mammograms digitized with these two different digitizers, the images digitized with LUMISCAN 85 digitizer were convolved with a 2×2 box filter and subsampled by a factor of two, resulting in 100 μm images.

In order to validate the prediction abilities of the classifier, the data set was partitioned randomly into training and test subsets. Approximately 73% of the samples have been used for training and 27% for testing. The data set was repartitioned randomly ten times and the training and test results were averaged to reduce their variability.

4.2. Feature extraction

The texture features used in this study were calculated from spatial grey-level dependence (SGLD) matrices[6,7,18] and run-length statistics (RLS) matrices[19]. The SGLD and RLS matrices were computed from the images obtained by the rubber band straightening transform (RBST)[8]. The RBST maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the mass border appears approximately as a horizontal edge, and spiculations appear approximately as vertical lines. A complete description of the RBST can be found in the literature[8].

The (i,j)th element of the SGLD matrix is the joint probability that gray levels i and j occur in a direction θ at a distance of d pixels apart in an image. Based on our previous studies[6], a bit depth of eight was used in the SGLD matrix construction, i.e., the four least significant bits of the 12 bit pixel values were discarded. Thirteen texture measures including correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, inertia, sum variance, difference average, difference variance and two types of information measure of correlation were used. These measures were extracted from each SGLD matrix at ten different pixel pair distances (d=1, 2, 3, 4, 6, 8, 10, 12, 16 and 20) and in four directions (0°, 45°, 90°; and 135°). Therefore, a total of 520 SGLD features were calculated for each image. The definitions of the texture measures are given in the literature[6-8,18]. These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. A gray level run is a set of consecutive, collinear pixels in a given direction which have the same gray level value. The run length is the number of pixels in a run[19]. The RLS matrix describes the run length statistics for each gray level in the image. The (i,j)th element of the RLS matrix is the number of times that the gray level i in the image possesses a run length of j in a given direction. In our previous study, it was found experimentally that a bit depth of 5 in the RLS matrix computation could provide good texture characteristics[8].

Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity, and run percentage were extracted from the vertical and horizontal gradient images in two directions, θ = 0°, and θ =90°. Therefore, a total of 20 RLS features were calculated for each ROI.

A total of 540 features (520 SGLD and 20 RLS) were therefore extracted from each ROI.

4.3. Feature selection

In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with stepwise linear discriminant analysis[20] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda was used as a selection criterion.

4.4. Performance analysis

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program[21], which fit a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. The discriminant scores of all case samples classified in the two stages of ART2LDA are combined. All masses classified into the malignant group by the ART2 stage were assigned a constant positive discriminant score higher than or equal to the most malignant discriminant score obtained from the LDA classifier.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve ($A_z^{(0.9)}$) at a true positive fraction (TPF) higher than 0.9. The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high sensitivity (low false negative) region which is most important for cancer detection in clinical practice.

5. RESULTS

In this study, the test subset was kept truly independent from the training subset; only the training subset was used for feature selection and classifier training, and only the test subset was used for classifier validation. In order to validate the prediction abilities of the classifier, ten different partitions of the training and test sets were used and the average classification results were estimated.

Table 1. Number of selected features for the 10 data groups.

Data Group No.	1	2	3	4	5	6	7	8	9	10	Mean
Number of selected features	12	15	13	18	14	14	13	18	14	14	14

For a given partition of training and test sets, feature selection was performed based on the training set. The feature selection results for the ten different training groups are shown in Table 1. The average number of selected features was 14. The selected feature sets contained an average of two RLS features and twelve SGLD features. A different ART2LDA classifier was trained using each training set and the corresponding set of selected features.

5.1. ART2LDA classification results

For the ART2LDA classifier, the number of selected features determines the dimensionality of the input vector of the ART2 classifier and the dimensionality of the LDA classifier. By using different values for the vigilance parameter, ART2 classifiers with different number of classes were obtained. In this study, the vigilance parameter p_{vig} was varied from 0.9 to 0.99, resulting in a range of 10 to 240 classes. The overall performance of the ART2LDA classifier was evaluated for different numbers of ART2 classes because different subset of the samples were separated and classified by ART2. In Fig. 5, the classification results for the ART2LDA are compared to the results from LDA alone for the training and test set partition no. 3. The classification accuracy, A_z, was plotted as a function of the number of ART2 classes. For this training and test set partition, when the number of classes was between 20 and 60, the ART2LDA classifier improved the classification accuracy for the test set in comparison to LDA. As the number of classes increased to greater than 60, the A_z value increased for the training data set, but decreased for the test data set and was lower than that of the LDA alone.

In Table 2 the A_z values of the test set for the 10 corresponding partitions are shown. The average test A_z value is 0.81 for the ART2LDA and 0.78 for LDA alone. For nine of the ten partitions, the A_z value was improved by the hybrid classifier.

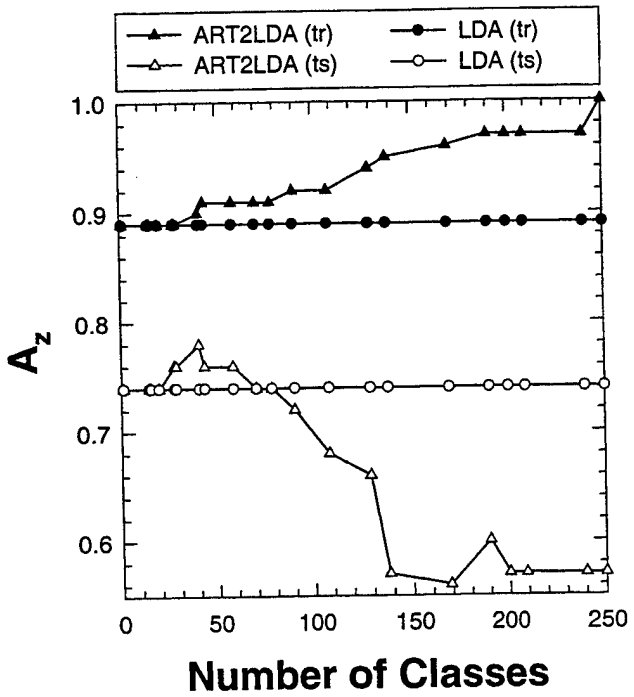


Figure 5. ART2LDA and LDA classification results for training and test sets from data group No.3 as a function of the number of classes generated by ART2.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve $A_z^{(0.9)}$ at a TPF higher than 0.9. In Table 3 the $A_z^{(0.9)}$ values of the test set for the 10 partitions of training and test sets are presented. The average test $A_z^{(0.9)}$ value is 0.34 for the ART2LDA and 0.27 for LDA. For nine of the ten partitions, the $A_z^{(0.9)}$ value was improved at the high sensitivity operating region (TPF>0.9) of the ROC curve.

Table 2. Classifiers performance for the 10 test sets. The A_z values represent the total area under ROC curve.

Data Group No.	LDA	ART2LDA
1	0.77	0.83
2	0.78	0.80
3	0.74	0.78
4	0.77	0.77
5	0.77	0.78
6	0.80	0.83
7	0.80	0.81
8	0.77	0.80
9	0.77	0.80
10	0.86	0.89
Mean	0.78	0.81

Table 3. Classifiers results for the 10 test sets. The A_z values represent the partial area of the ROC curve above the true positive fraction of 0.9 ($A_z^{(0.9)}$).

Data Group No.	LDA	ART2LDA
1	0.14	0.23
2	0.17	0.21
3	0.19	0.32
4	0.19	0.21
5	0.24	0.26
6	0.27	0.38
7	0.32	0.31
8	0.32	0.34
9	0.40	0.49
10	0.44	0.60
Mean	0.27	0.34

6. DISCUSSION

In this paper a new classifier (ART2LDA) is designed and applied to the classification of malignant and benign masses. The results indicate that the ART2LDA classifier has better generalizability than an LDA classifier alone. The ART2 classifier groups the case samples that are different from the main population into separate classes. The minimum number of classes needed to start the clustering of outliers into separate classes depends on how different the outliers are from the rest of the sample population. For the ten different partitions of the training and test sets used in this study, the minimum number varied between 13 and 15 classes. When the number of ART2 classes was less than this minimum number of classes, the ART2 classifier generated only mixed malignant-benign classes and all samples were transferred to the LDA stage. In that case, the ART2LDA was equivalent to the LDA classifier alone. When a higher number of classes was generated, an increased number of cases that may be considered outliers of the general data population was removed (clustered in separate classes). For the ten training sets used in this study, the malignant outliers were gradually removed when the number of classes increased. The training accuracy increased when the number of classes increased and A_z could reach the value of 1.0. However, a large number of ART2 classes led to overfitting the training sample set and poor generalization in the test set. The classification accuracy of ART2 for the test set tended to decrease when the number of classes was greater than about 70. The large number of classes also led to a reduction in the generalizability of the second-stage LDA; the training of LDA with a small number of samples would again result in overfitting the training set, and poor generalizability in the test set. This effect was observed when more than 60 or 70 classes were generated by ART2 (see Fig. 5).

The classification accuracy of ART2LDA increased initially with increased number of classes and then decreased after reaching a maximum. The correct classification of the outliers by the ART2 in combination with an improvement in the classification by the LDA resulted in the increased accuracy. When the number of ART2 classes was further increased, the effects of overfitting by the ART2 and the LDA became dominant and the prediction ability of the ART2LDA decreased. In some cases the second stage LDA prediction was much worse than the ART2. In other cases the ART2 could not generalize well. The generation of a high number of classes is therefore impractical and unnecessary both from computational and methodological point of view.

When the partial area of the ROC curve above the true positive (TP) fraction of 0.9 ($A_z^{(0.9)}$) was considered as a measure of classification accuracy, the advantage of ART2LDA over LDA alone became even more evident. By removing and correctly classifying the outliers the accuracy of the classification is increased at the high sensitivity end of the curve.

We have performed statistical tests with the CLABROC program to estimate the significance in the differences between the A_z values from the ART2LDA and the LDA alone, as well as in the differences in the partial $A_z^{(0.9)}$ from the two classifiers. The statistical tests were performed for each individual data set partition because the correlation among the data sets from the different partitions precludes the use of Student's paired t-test with the ten partitions. We found that the differences in both cases did not reach statistical significance because of the small number of test samples and thus the large standard deviation in the A_z values. However, the consistent improvements in A_z and $A_z^{(0.9)}$ (9 out of 10 data set partitions in both cases) suggest that the improvement was not by chance alone, and that the accuracy of a classification task could be improved by the use of an ART2 network.

An important difference between the classifier designed in this study and many others in the CAD field is the method of feature selection. In several previously published studies[8,22,23] the features were selected from the entire data set first, and then the data set was partitioned into training and test sets. This meant that at the feature selection stage of the classifier design, the entire data set was considered to be a training set. Depending on the distribution of the features and the total number of samples used, the test results in these studies might be optimistically biased[24]. In this study, initially the entire data set was partitioned into training and test sets and then feature selection was performed only on the training set. This method results in a pessimistic estimate of the classifier performance[24] when the training set is small. We therefore expect that the performance will be improved when the classifier designed in this study is trained using a large data set. Since our main purpose in this study was to compare the LDA and ART2LDA classifiers, we did not attempt to quantify how pessimistic our results are in this study.

7. CONCLUSION

A new classifier combining an unsupervised ART2 and a supervised LDA has been designed and applied to the classification of malignant and benign masses. A data set consisting of 348 films (179 malignant and 169 benign) was

randomly partitioned into training and test subsets. Ten different random partitions were generated. For each training set, texture features were extracted and feature selection was performed. An average of fourteen features were selected for each group. Ten hybrid ART2LDA classifiers and ten LDA models alone were trained by using the ten training sets. The average A_z value under the ROC curve for the test sets was better for ART2LDA (A_z=0.81) compared to the LDA alone (A_z=0.78). A greater improvement was obtained when the partial ROC area above a true-positive fraction of 0.9 was considered. The average partial A_z for ART2LDA was 0.34 as compared to 0.27 for LDA. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

ACKNOWLEDGMENTS

This work was supported by a USPHS Grant No. CA 48129, and by U.S. Army Medical Research and Materiel Command (USAMRMC) Grant DAMD 17-96-1-6254. Lubomir Hadjiiski is also supported by a Career Development Award from the USAMRMC (DAMD 17-98-1-8211). Berkman Sahiner is also supported by a Career Development Award from the USAMRMC (DAMD 17-96-1-6012). Nicholas Petrick is also supported by a grant from The Whitaker Foundation. The content of this publication does not necessary reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. We would like to thank Prof. Stephen Grosberg and Dr. Gail Carpenter for providing us with valuable information as well as for the useful discussions. Additionally we would like to thank Charles E. Metz, Ph.D. for providing the LABROC1 and CLABROC programs.

REFERENCES

1. H. C. Zuckerman, "The role of mammography in the diagnosis of breast canser," in *Breast Canser, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw-Hill, New York, 1987), pp. 152-172.
2. D. B. Kopans, "The positive predictive value of mammography," *Am. J. Roentgenol.* 158, pp. 521-526, 1992.
3. D. D. Adler, and M. A. Helvie, "Mammographic biopsy recommendations," *Curr. Opin. Radiol.* 4, pp. 123-129, 1992.
4. R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York), 1973.
5. D. Rumelhart, G. E. Hinton, and R. J. Williams, in D. E. Rumelhart (ed.), *Parallel and Distributed Processing*, Vol. 1, MIT Press, 1986, pp. 318.
6. H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer- Aided Classification of Mammographic Masses and Normal Tissue: Linear Discriminat Analysis in Texture Feature Space," *Phys. Med. Biol.* 40, pp. 857-876, 1995.
7. D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of Mass and Normal Breast Tissue on Digital Mammograms: Multiresolution Texture Analysis," *Med. Phys.*, 22, pp. 1501-1513, 1995.
8. B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, "Computerized Characterization of Masses on Mamograms: The Rubber Band Sraightening Transform and Texture Analysis," *Med. Phys.* 25 (4), pp. 516-526, April 1998.
9. H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler and M. M. Goodsitt, "Computerized Classification of Malignant and Benign Microcalsifications on mammograms: Texture analysis using an Artificial Neural Network," *Phys. Med. Biol.* 42, pp. 549-567, 1997.
10. M. Jordan, and R. A. Jacobs, "Hierarchical Mixture of Experts and EM Algorithm," *Neural Computation*, 6, pp. 181-214, 1994.
11. L. Hadjiiski, and P. Hopke, "Design of Large Scale Models Based on Multiple Neural Network Approach," *Intelligent Engineering Systems Through Artificial Neural Networks*, Vol. 7, ASME Press, 1997, pp. 61-66.
12. S. Grossberg, "Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors," *Biological Cybernetics*, vol.23, no.3, pp.121-134, 1976.
13. S. Grossberg, "Adaptive pattern classification and universal recoding. II. Feedback, expectation, olfaction, illusions," *Biological Cybernetics*, vol.23, no.4, pp. 187-202, 1976.
14. G. A. Carpenter, and S. Grossberg, "ART 2: self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol.26, no.23, 1, pp. 4919-4930, Dec. 1987.
15. G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition," *Neural-Networks*, vol.4, no.4, pp. 493-504, 1991.
16. G. A. Carpenter, and N. Markuzon, "ARTMAP-IC and Medical Diagnosis: Instance Counting and Inconsistent Cases," *Neural-Networks*, vol.11, no.2, pp. 323-336, March 1998.

17. Y. Xie, P. K. Hopke, and D. Wienke, "Airborne Particle Classification with a Combination of Chemical Composition and Shape Index Utilizing an Adaptive Resonance Artificial Neural network," *Environmental Science & Technology*, Vol. 28, No. 11, pp. 1921-1928, 1994.
18. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst. Man Cybern.* 3, pp. 610-621, 1973.
19. M. M. Galloway, "Texture Analysis Using Gray Level Run Length," *Comput. Graph. Image Process.* 4, pp. 172-179, 1975.
20. M. J. Norusis, SPSS Professional Statistics 6.1 (SPSS Inc., Chicago, 1993).
21. C. E. Metz, J. H. Shen, and B. A. Herman, "New Methods for Estimating a Binomial ROC Curve From Continuously Distributed Test Results," *presented at the 1990 Annual Meeting of the American Statistical Association, Anahaim, CA*, 1990.
22. M. F. McNitt-Gray, H. K. Huang, J. W. Sayre, "Feature Selection in the Pattern Classification Problem of Digital Chest Radiograph Segmentation," *IEEE Transaction on Medical Imaging*, Vol. 14, No. 3, pp. 537-547, Sep. 1995.
23. Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms," *Acad. Radiol.*, 5, pp. 155-168, 1998.
24. B. Sahiner, H. P. Chan, N. Petrick, R. Wagner, L. Hadjiiski, "The effect of sample size on feature selection in computer-aided diagnosis," *SPIE International Symposium on Medical Imaging*, San Diego, California, February 20-26, 1999., *Proc. SPIE* 3661, (in print).

ACTIVE CONTOUR MODELS FOR SEGMENTATION AND CHARACTERIZATION OF MAMMOGRAPHIC MASSES

Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick,

Lubomir M. Hadjiiski, Mark A. Helvie, Sophie Paquerault

Department of Radiology, University of Michigan, Ann Arbor, MI 48109

Abstract

We have investigated the use of an active contour model for accurate delineation of mass boundaries on mammograms. The model used smoothness constraints and image gradient information in order to refine an initial boundary provided by a clustering algorithm. After segmentation of the mass, possible spiculations were segmented by utilizing gradient direction statistics in a region surrounding the mass. Spiculation measures and morphological features were extracted and used for classifying the mass as malignant or benign. The classification accuracy was evaluated using the area A_z under the receiver operating characteristic (ROC) curve. A data set containing 243 mammograms from 101 patients was used for training the classifier, and a data set containing 95 mammograms from 45 patients were used for testing the classifier. The test A_z for the task of classifying a mass on a single view and a mass on all available views as malignant or benign was 0.81 and 0.87, respectively. Our results indicate that the spiculation measures and the morphological features extracted from automatically segmented mass boundaries are effective in characterizing mammographic masses as malignant or benign.

1. Introduction

In recent years, many researchers have investigated the use of computer-extracted image features for classification of breast masses as malignant or benign (Sahiner *et al.* 1998; Huo *et al.* 1998; Leichter *et al.* 2000). Many features used in computerized breast mass characterization require accurate delineation of mass boundaries as a first step. Accurate computerized delineation of

mass boundaries is often difficult because of the presence of ill-defined or obscured boundaries. The human visual system often overcomes this problem by incorporating *a-priori* information, such as smoothness of mass boundaries, with the image information. In order to make use of similar information for computerized mass segmentation, we designed an active contour model based on the image characteristics of mammographic masses. The new model was used to improve the boundaries provided by a clustering algorithm that was developed in our earlier studies. After segmentation, morphological features were extracted from the mass shape, and were combined with spiculation measures for the characterization of breast masses as malignant or benign.

2. Mass segmentation

The location of the biopsied mass was identified by an MQSA-approved radiologist. A region of interest (ROI) containing the biopsied mass was extracted from the mammogram for computerized processing.

2.1. Initial mass segmentation

The mass segmentation method employed in this study started with the initial detection of a mass shape within an ROI using a K-means clustering algorithm. This technique has been discussed in detail in the literature (Sahiner *et al.* 1996). Figures 1(a)-(d) show examples of a spiculated and a nonspiculated mass, and the results of the initial segmentation.

2.2. Active contour segmentation

Although clustering-based mass segmentation resulted in reasonable mass shapes for most of the masses, the segmentation exhibited inaccuracies when the mass was not very conspicuous, or when some parts of the mass were obscured by overlapping normal breast structures. In addition, further refinement was necessary before detection and segmentation of spiculations.

We used an active contour model for the first stage mass shape refinement, and spiculation detection and segmentation for the final shape refinement.

An active contour is a deformable continuous curve, whose shape is controlled by internal forces (the model, or *a-priori* knowledge about the object to be segmented) and external forces (the image). The internal forces impose a smoothness constraint on the contour, and the external forces push the contour towards salient image features, such as edges. To solve a segmentation problem, an initial boundary is iteratively deformed so that the energy due to internal and external forces is minimized along the contour.

The internal energy components in our active contour model were the continuity and curvature of the contour, as well as the homogeneity of the segmented object. The external energy components were the negative of the smoothed image gradient magnitude, and a balloon force that exerted pressure at a normal direction to the contour. The contour was represented by the vertices of an N-point polygon whose vertices were $\mathbf{v}(i)=(x(i),y(i))$, $i=1,\ldots,N$. The energy to be minimized was defined as

$$E=\sum_{i=1}^{N}\left[w_{curv}E_{curv}(i)+w_{cont}E_{cont}(i)+w_{grad}E_{grad}(i)+w_{bal}E_{bal}(i)\right]+w_{\text{hom}}E_{\text{hom}} \qquad (1)$$

where each energy term has a weight, w.

The curvature energy term is represented by an approximation to the second derivative of the contour, $E_{curv}(i)=\left|\mathbf{v}(i-1)-2\mathbf{v}(i)+\mathbf{v}(i+1)\right|$. This term is large when the angle at vertex i is small. By discouraging small angles at vertices, this term attempts to smooth the contour. The continuity term, $w_{cont}E_{cont}(i)$, reflects the deviation of the length of the line segment under

consideration from the average line segment length $\bar{d}$. This term favors contours with regular spacing between the vertices over those with irregular spacing. The image gradient magnitude is obtained by smoothing the image with a low-pass filter, finding the partial derivatives in the horizontal and vertical directions, and then computing the magnitude of the partial derivative vector. Since the gradient energy, $E_{grad}(i)$, is defined as the negative of the gradient magnitude, minimizing this term attracts the contour to object edges. The balloon energy encourages the contour to expand in the normal direction, which is required to prevent the contour from collapsing onto itself (Cohen 1991). The purpose of the homogeneity term, $w_{hom}E_{hom}(i)$, is to make the object and the background regions as homogeneous as possible within each region, and to maximize the difference between the two regions (Poon and Braun 1997).

To minimize the contour energy, we used a greedy algorithm that was first proposed by Williams and Shah (Williams and Shah 1992). In this algorithm, the contour was iteratively optimized, starting with the initial contour provided by clustering-based segmentation. At each iteration, a neighborhood of each vertex was examined, and the vertex was moved to the location that minimized the contour energy. Figures 1(c)-(f) show the initial and final contours, respectively, of the model for a spiculated and a nonspiculated mass.

2.3. Segmentation of spiculations

Spiculations on mammograms appear as linear structures with a positive image contrast, and they usually lie in a radial direction to the mass. As a result of their linearity, the gradient directions at image pixels on or close to the spiculation are more or less in the same orientation relative to that of the spiculation. In order to investigate whether a pixel (i_c, j_c) on the mass contour lies on the path of a spiculation, one can make use of this property as follows: In a search region S of the image, compute the statistics of the angular difference θ between the image gradient direction

at image pixel (i,j), and the direction of the vector joining pixels (i_c,j_c), and (i,j) (figure 2). If the pixel (i_c,j_c) lies on the path of a spiculation, then θ will be close to $\pi/2$ whenever the image pixel (i,j) is on the spiculation. Therefore, the distribution of θ, obtained from all image pixels (i,j) within the search region S will have a peak around $\pi/2$. If there is no spiculation, and if the gray levels in S are randomly distributed, then this distribution will be uniform. Karssemeijer *et al.* have made use of a similar idea for detecting spiculated lesions on mammograms (Karssemeijer and te Brake 1996), but not for the detection of the actual spiculations. In our method, we combined this idea with the fact that spiculations generally lie in a radial direction to the mass. Therefore, the region S could be limited so that other gradients, such as those resulting from the mass contour itself, can be excluded from the distribution of gradients in S. The details of our spiculation detection method are described in the literature (Sahiner *et al.* 2000; Chan *et al.* 2000). The contours of a spiculated and a nonspiculated mass after spiculation detection are shown figures 1(g) and 1(h), respectively.

3. Feature Extraction and Classification

In the spiculation segmentation stage, three spiculation measures were extracted from each ROI. These were the number of possible spiculations (NPS), the percentage area of spiculations (PAS), and the product of these two measures (PR). These spiculation measures were used in addition to eleven morphological features extracted from the final mass outline for mass characterization. The first five morphological features were based on the normalized radial length (NRL), defined as the Euclidean distance from the object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object. These features included NRL mean, standard deviation, entropy, area ratio, and zero crossing count (Petrick *et al.* 1999). The remaining six morphological features included the perimeter, area, perimeter-to-

area ratio, circularity, rectangularity, and contrast of the object. The definition of these features can be found in the literature (Petrick *et al.* 1999).

Stepwise feature selection was used to select effective features for classification from the feature space of fourteen features. Four features, namely, NPS, PR, contrast, and circularity were selected using the set of training ROIs. A backpropagation neural network (BPN) with four input nodes, two hidden-layer nodes, and a single output node was trained using the training set. The accuracy of the designed classifier was evaluated by applying the classifier to test cases that had not been used for training. The test scores were analyzed using receiver operating characteristic (ROC) methodology. The classification accuracy was evaluated as the area A_z under the ROC curve.

4. Data Set

The mammograms used in this study were randomly selected from the files of patients in the Radiology Department at the University of Michigan who had undergone biopsy. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass, and that approximately equal numbers of malignant and benign masses were present in the data set. Our training data set consisted of 243 mammograms (116 benign and 127 malignant) from 101 patients. Our test data set consisted of 95 mammograms (42 benign and 53 malignant) from 45 patients. A single view was available for nine of these 45 patients. For the remaining 36 test patients, two or more views were available. The true pathology of all the masses was determined by biopsy and histologic analysis.

5. Results

We investigated film-based classification of the masses on each mammogram, as well as case-based classification by combining possible multiple views of the same mass. For case-based

classification, the BPN scores from different views were averaged. The training A_z values for film-based and case-based classification were 0.91 and 0.95 respectively. The test A_z values for film-based and case-based classification were 0.81 and 0.87. The training and test ROC curves are shown in figures 3(a) and 3(b), respectively.

6. Discussion and Conclusion

In our previous work, the clustering method was successful in segmenting the main portion of the mass from the background. However, a major limitation of clustering-based segmentation is that, even for well-circumscribed masses, the segmented shape contains many irregularities due to structured or random noises (see figure 1(d)). Another limitation is that, when parts of the mass are obscured by overlapping normal breast structure, clustering method yields inaccurate results. In this study, we used an active contour model for refining the clustering-based segmentation results. By choosing a balance between the active contour weights based on the training set, we were able to obtain object shapes that were mostly smooth, but contours with sharp turns were also possible if the object boundary contained large gradients. Compared to clustering, the resulting boundaries were subjectively judged to be closer to actual mass boundaries. However, the active contour model was not suitable for the segmentation of spiculations. Since the spiculations do not have a large gradient magnitude, the contour cannot have sharp turns at spiculation locations unless w_{curv} is very small. However, a small value for w_{curv} is not practical, because it results in mass shapes that are too irregular all around the contour. For this reason, we designed an additional stage for detection and segmentation of spiculations.

Our results indicate that accurate segmentation of mammographic masses, detection of spiculations, and the use of morphological and spiculation features can be effective in classifying breast masses as malignant or benign.

Acknowledgments

This work is supported by USPHS Grant CA 48129, by a Career Development Award (B.S.) from the USAMRMC (DAMD 17-96-1-6012), and a Whitaker Foundation Grant (N.P.). The content of this publication does not necessarily reflect the position of the funding agencies, and no official endorsement of any equipment and product of any companies mentioned in this publication should be inferred.

References

Chan, H.-P., N. Petrick and B. Sahiner (2000). Computer-aided breast cancer diagnosis. *Soft Computing Techniques in Breast Cancer Prognosis and Diagnosis* Ed. L. C. Jain. New York, CRC Press. (in press).

Cohen, L. D. 1991. On active contour models and baloons. *CVGIP: Img. Underst.* 53: 211-218.

Huo, Z. M., M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt and K. Doi. 1998. Automated computerized classification of malignant and benign masses on digitized mammograms. *Acad. Rad.* 5: 155-168.

Karssemeijer, N. and G. te Brake. 1996. Detection of stellate distortions in mammograms. *IEEE Trans. Med. Img.* 15(5): 611-619.

Leichter, I., S. Fields, R. Nirel, P. Bamberger, B. Novak, R. Lederman and S. Buchbinder. 2000. Improved mammographic interpretation of masses using computer-aided diagnosis. *Eur. Radiol.* 10: 377-383.

Petrick, N., H. P. Chan, B. Sahiner and M. A. Helvie. 1999. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Med. Phys.* 26(8): 1642-1654.

Poon, C. S. and M. Braun. 1997. Image segmentation by a deformable contour model incorporating region analysis. *Phys. Med. Biol.* 42: 1833-1841.

Sahiner, B., H. P. Chan, N. Petrick, M. A. Helvie and M. M. Goodsitt. 1998. Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. *Med. Phys.* 25: 516-526.

Sahiner, B., H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler and M. M. Goodsitt. 1996. Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms. *Med. Phys.* 23: 1671-1684.

Sahiner, B., H.-P. Chan, N. Petrick, M. A. Helvie and L. M. Hadjiiski. 2000. Improvement of mammographic mass characterization using spiculation measures and morphological features. *Med. Phys.* (submitted):

Williams, D. J. and M. Shah. 1992. A fast algorithm for active contours and curvature estimation. *CVGIP: Img. Underst.* 55: 14-26.

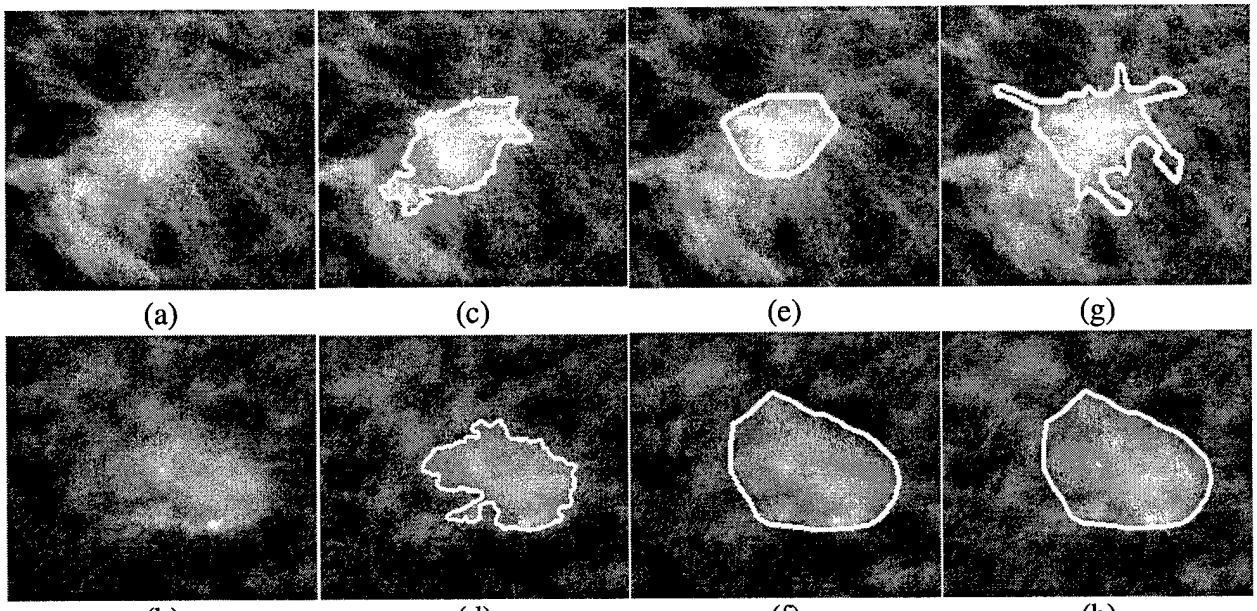


Figure 1. (a), (b) The mass ROI, (c), (d) clustering-based segmentation, (e), (f) active-contour based segmentation, and (g), (h) the result of spiculation detection and segmentation for a spiculated mass (a, c, e, and g) and a nonspiculated mass (b, d, f, and h).

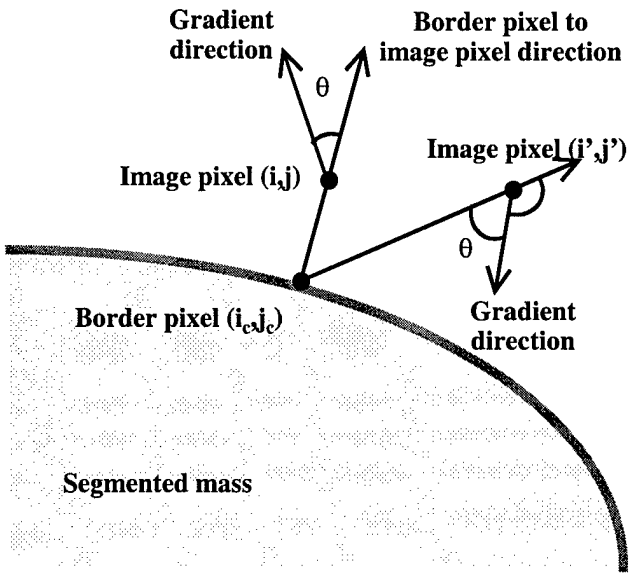


Figure 2. The definition of the angular difference θ.

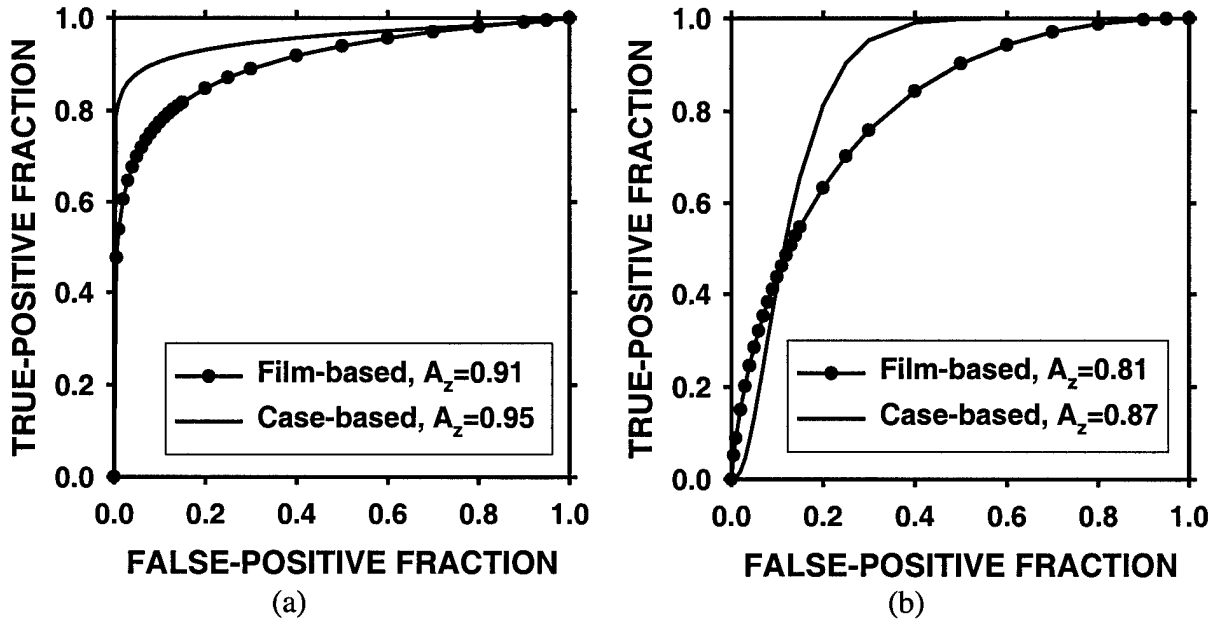


Figure 3. ROC curves for film-based and case-based classification. (a) Training (b) Test.

Analysis of temporal change of mammographic features for computer-aided characterization of malignant and benign masses

Lubomir Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Mark A. Helvie, Metin Gurcan

Department of Radiology, The University of Michigan, Ann Arbor, MI 48109-0904

ABSTRACT

A new classification scheme was developed to classify mammographic masses as malignant and benign by using interval change information. The masses on both the current and the prior mammograms were automatically segmented using an active contour method. From each mass, 20 run length statistics (RLS) texture features, 3 spiculation features, and mass size were extracted. Additionally, 20 difference RLS features were obtained by subtracting the prior RLS features from the corresponding current RLS features. The feature space consisted of the current RLS features, the difference RLS features, the current and prior spiculation features, and the current and prior mass sizes. Stepwise feature selection and linear discriminant analysis classification (LDA) were used to select and merge the most useful features. A leave-one-case-out resampling scheme was applied to train and test the classifier using 140 temporal image pairs (85 malignant, 55 benign) obtained from 57 biopsy-proven masses (33 malignant, 24 benign) in 56 patients. An average of 10 features were selected from the 56 training subsets: 4 difference RLS features, 4 RLS features and 1 spiculation feature from the current image, and 1 spiculation feature from the prior, were most often chosen. The classifier achieved an average training A_z of 0.92 and a test A_z of 0.88. For comparison, a classifier was trained and tested using features extracted from the 120 current single images. This classifier achieved an average training A_z of 0.90 and a test A_z of 0.82. The information on the prior image significantly ($p=0.01$) improved the accuracy for classification of the masses.

Keywords: Computer-Aided Diagnosis, Interval Changes, Classification, Feature analysis, Mammography, Malignancy.

1. INTRODUCTION

Mammography is currently the most effective method for early breast cancer detection[1,2]. Analysis of interval changes is an important method used by radiologists in mammographic interpretation to detect developing malignancy[3,4]. A variety of computer-aided diagnosis (CAD) techniques have been developed to detect mammographic abnormalities and to distinguish between malignant and benign lesions. We are studying the use of CAD techniques to assist radiologists in interval change analysis.

Commonly used classification methods for CAD use information from a single image. These methods have been shown to perform well in lesion classification problems[6-13]. However, when multiple-year mammograms of a mass are available, it is not trivial to design computer vision methods to use the temporal information for computer-aided classification and to improve the differentiation between benign and malignant masses.

The goal of our research is to develop a technique for computerized analysis of temporal differences between a lesion on the most recent mammogram and a prior mammogram of the same view. The computer algorithm can be used to assist radiologists in evaluating interval changes and thus distinguishing between malignant and benign masses for CAD. We have previously presented[5] preliminary results that demonstrated the feasibility of classifying malignant and benign masses based on interval change analysis. In this study, we continue the development of this approach. Additionally, we compared this method with a classification method based on information extracted from the current mammogram alone.

2. CLASSIFICATION TECHNIQUE

A new classification scheme was developed to classify mammographic masses as malignant and benign by using interval change information. The technique is based on the generation of features that we expect will represent adequately the temporal information and will discriminate between malignant and benign masses.

In *Medical Imaging 2001*: Image processing, Milan Sonka, Kenneth M. Hanson, Editors,
Proceedings of SPIE Vol. 4322 (2001)

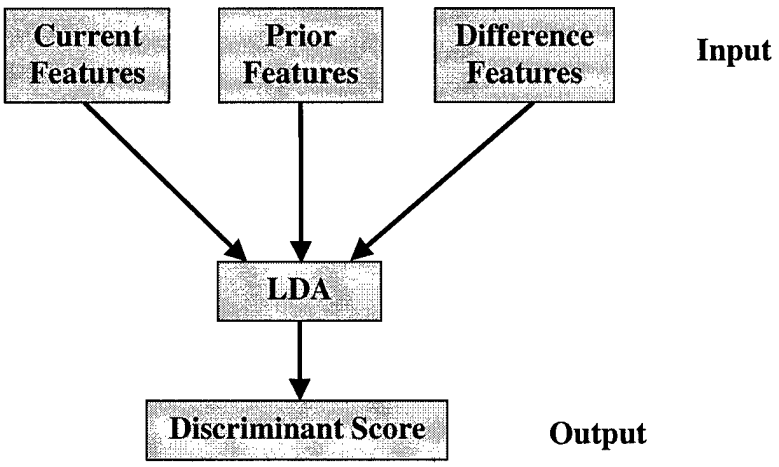


Figure 1. Block-diagram of the classification method.

The mass to be analyzed can either be identified manually by a radiologist or automatically by a computerized detection program. In this study, the masses were identified by an MQSA radiologist on each mammogram. The masses on both the current and the prior mammograms were automatically segmented using an active contour method. An example of the segmentation is shown in Figure 2 and Figure 3 for a malignant and a benign mass, respectively. Features such as texture features, spiculation features and mass size were extracted from each mass. Additionally, difference features were obtained by subtracting a prior feature from the corresponding current feature. The feature space consisted of current, prior, and difference features. Stepwise feature selection applied to linear discriminant analysis (LDA) were used to select the most useful features. The selected features were then used as the input predictor variables of the LDA classifier (Figure 1). A leave-one-case-out resampling scheme was employed to train and test the classifier. The LDA classifier was used in order to keep the discrimination function simple, thereby reducing the possibility of over-training.

To evaluate the improvement in the classifier performance designed by using the temporal change information, an additional classifier was trained using the information extracted from the current single images of the temporal pairs. We will refer to these images as current images. Comparison of the two classifiers will reveal the effectiveness of interval change analysis on classification of malignant and benign masses.

3. DATA SET

A set of 140 temporal pairs of mammograms containing biopsy-proven masses on the current mammograms was used to examine the performance of this approach. The data set consisted of a total of 241 mammograms from 56 patients. The mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of $50\,\mu m \times 50\,\mu m$ and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly proportional to the optical density (OD) within the range of 0 to 4 OD units, with a slope of 0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The digitizer output was linearly converted so that a large pixel value corresponded to a low optical density. The images were averaged and down-sampled by a factor of 2 resulting in images with a pixel size of $100\,\mu m \times 100\,\mu m$ for further analysis.

The 56 cases contained 57 biopsy proven masses (33 malignant and 24 benign). The 241 mammograms contained different mammographic views and multiple years of the masses including the year when the biopsy was performed. By matching masses of the same view from two different exams, a total of 140 temporal pairs were formed, of which 85 were malignant and 55 benign. A malignant temporal pair consisted of a biopsy proven malignant mass or a mass that was followed up and found to be malignant by biopsy in a future year. Similar definitions were used for the benign temporal pairs. Within the 140 temporal pairs, a total of 120 mammograms were current mammograms. Of the 120 current mammograms, 70 were malignant and 50 benign.

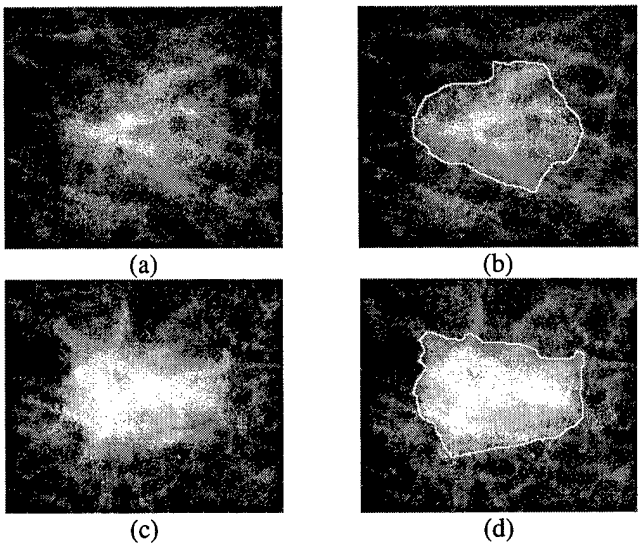


Figure 2. A malignant mass: (a) the mass in a prior year mammogram (1997), (b) mass outline obtained by active contour segmentation, (c) the mass in a current year mammogram (1998), (d) mass outline obtained by active contour segmentation.

Since all cases in this data set had undergone biopsy, the benign masses in this set could not be distinguished easily from the malignant ones based on current mammographic criteria. Examples of such cases are shown in Figure 2 and Figure 3. The malignant mass in Figure 2 did not increase in size but changed its density. The benign mass (Figure 3), on the other hand, appeared to have spicules. For the malignant masses in this data set, the average mass size was 8.2 mm on the prior mammograms and 12.7 mm on the current mammograms. The corresponding sizes were 10.6 mm and 12.2 mm, respectively, for the benign masses. The temporal pairs had a time interval of 6 to 36 months. More than 70% of the pairs had a time interval of 12 months.

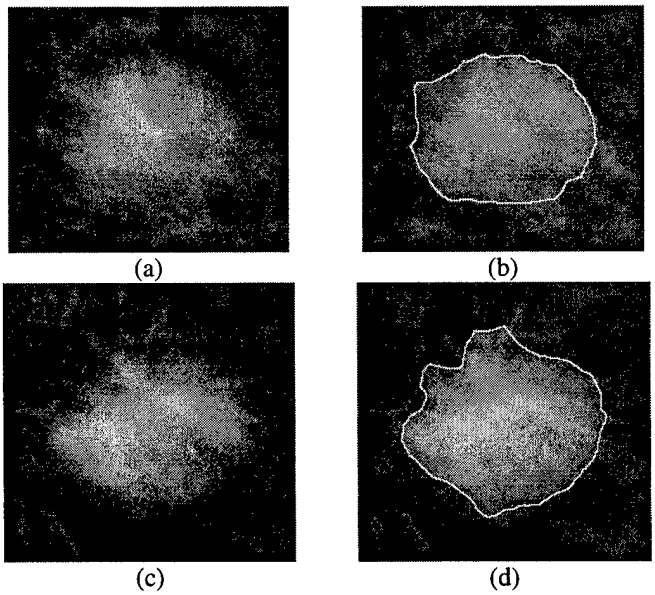


Figure 3. A benign mass: (a) the mass on a prior year mammogram (1995), (b) mass outline obtained by active contour segmentation, (c) the mass on a current year mammogram (1996), (d) mass outline obtained by active contour segmentation.

4. FEATURE EXTRACTION

A rectangular region of interest (ROI) was defined to include the radiologist-identified mass with an additional surrounding breast tissue region of at least 40 pixels wide from any point of the mass border. A fully automated method was then used for segmentation of the mass from the breast tissue background within the ROI. The masses on both the current and the prior mammograms were automatically segmented using a 2D active contour method, initialized by adaptive thresholding[14,15].

The texture features used in this study were calculated from run-length statistics (RLS) matrices[18]. The RLS matrices were computed from the images obtained by the rubber band straightening transform (RBST)[7]. The RBST maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the mass border appears approximately as a horizontal edge, and spiculations appear approximately as vertical lines. A complete description of the RBST can be found in the literature[7].

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity, and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta = 0^\circ$, and $\theta = 90^\circ$. Therefore, a total of 20 RLS features were calculated for each ROI. The definition of the RLS feature measures can be found in the literature[18].

The morphological features were extracted from the automatically segmented mass shape, and included features such as the area, circularity, rectangularity, compactness, and the axis ratio[15]. Spiculation features were extracted by using the statistics of the image gradient direction relative to the normal direction to the mass border in a ring of pixels surrounding the mass[14,15].

A total of 35 features (20 RLS, 12 morphological and 3 spiculation) were therefore extracted from each ROI. Additionally, difference features were obtained by subtracting a prior feature from the corresponding current feature. Therefore 20 RLS, 12 morphological and 3 spiculation difference features were obtained.

5. FEATURE SELECTION

In order to reduce the number of the features and to obtain the best feature subset to design an effective classifier, feature selection with stepwise linear discriminant analysis[19,20] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool based on analysis of its effect on the selection criterion. The stepwise selection procedure is controlled by a simplex optimization method[16, 17] in such a way that a minimum number of features were selected to achieve a high accuracy of classification by LDA. More details about the stepwise linear discriminant analysis and its application to CAD can be found elsewhere[6, 7].

6. EVALUATION METHODS

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology[21]. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program[22], which fits a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. The performances of the classifiers were also assessed by estimation of the partial area index ($A_z^{(0.9)}$). The partial area index ($A_z^{(0.9)}$) is defined as the area that lies under the ROC curve but above a sensitivity threshold of 0.9 ($TPF_0 = 0.9$) normalized to the total area above TPF_0, ($1\text{-}TPF_0$). The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high sensitivity (low false negative) region which is most important for a cancer detection task.

7. CLASSIFICATION RESULTS

For the data set used in this study, an average of 10 features were selected from the 56 training subsets. The most frequently selected features included 4 difference RLS features, 4 RLS features and 1 spiculation feature from the current image, and 1 spiculation feature from the prior. The LDA classifier achieved an average training A_z of 0.92 and a test A_z of

0.88. The LDA classifier using features extracted from the current single images (the current images of the temporal pairs) achieved an average training A_z of 0.90 and a test A_z of 0.82. An average of 11 features were selected from the 56 training subsets. The most frequently selected features were 4 RLS features, 1 spiculation feature from the current image and 6 morphological features. The difference in the test A_z between the two classifiers is statistically significant (p=0.01). The classifier based on temporal pairs achieved a test partial $A_z^{(0.9)}$ of 0.37 and the classifier based on current images achieved a test $A_z^{(0.9)}$ of 0.32. These results are summarized in Table 1.

Table 1. Classification results for the classifier based on the temporal change information and the classifier based on current single image information.

Classification	Avg. no. of selected features	Training A_z	Test A_z	Test partial $A_z^{(0.9)}$
Temporal pairs	10	0.92	0.88 ± 0.028	0.37 ± 0.1
Current images	11	0.90	0.82 ± 0.038	0.32 ± 0.08

8. CONCLUSION

The difference RLS texture features and spiculation features were useful for identification of malignancy in temporal pairs of mammograms. The information on the prior image was important for characterization of the masses; 5 out of the 10 selected features contained prior information. We found that the size of the mass was not a discriminatory feature for these difficult cases because many of the benign masses also grew over time. The temporal change information significantly (p=0.01) improved the accuracy for classification of the masses in terms of the total area under the ROC curve (A_z). The partial area under the ROC curve is also improved for the classifier based on current and prior images ($A_z^{(0.9)} = 0.37$) compared to the classifier based only on the current images ($A_z^{(0.9)} = 0.32$), although the difference did not achieve statistical significance. Further studies are underway to improve this temporal change classification technique and to evaluate its performance on a larger data set.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award from the USAMRMC (DAMD 17-98-1-8211) (L.H.), a USPHS Grant CA 48129, and a USAMRMC grant (DAMD 17-96-1-6254).

REFERENCES

1. H. C. Zuckerman, "The role of mammography in the diagnosis of breast cancer," in *Breast Cancer, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw-Hill, New York, 1987), pp. 152-172.

2. L. Tabar and P. B. Dean, "The control of breast cancer through mammographic screening: What is the evidence," *Radiol. Clin. N. Amer.* **25**, pp. 993-1005, 1987.

3. L. W. Bassett, B. Shayestehfar and I. Hirbawi, "Obtaining previous mammograms for comparison: usefulness and costs," *Amer. J. Roentgenology* **163**, pp. 1083-1086, 1994.

4. E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: results in 3183 consecutive cases," *Radiology* **179**, pp. 463-468, 1991.

5. L. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, M. Gurcan, "Computer-Aided Classification of Malignant and Benign Breast Masses by Analysis of Interval Change of Features in Temporal Pairs of Mammograms", *Radiology* 2000 **217(P)**: 435.

6. H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, Computer- Aided Classification of Mammographic Masses and Normal Tissue: Linear Discriminat Analysis in Texture Feature Space, *Phys. Med. Biol.* **40**, pp. 857-876, 1995.

7. B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, Computerized Characterization of Masses on Mamograms: The Rubber Band Straightening Transform and Texture Analysis, *Med. Phys.* **25** (4), pp. 516-526, April 1998.

8. H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler and M. M. Goodsitt, Computerized Classification of Malignant and Benign Microcalsifications on mammograms: Texture analysis using an Artificial Neural Network, *Phys. Med. Biol.* **42**, pp. 549-567, 1997.

9. L. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M. Helvie, Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach, *IEEE Transactions on Medical Imaging*, Vol. **18**, No. 12, Dec. 1999, pp 1178-1187.

10. Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C.E. Metz, Artificial Neural Networks in Mammography: Application to Decision Making in the Diagnosis of Breast Cancer, *Radiology* **187**, pp. 81-87, 1993.

11. V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, Improvements in Specificity of Ultrasonography for Diagnosis of Breast Tumors by Means of Artificial Intelligence, *Med. Phys.*, **19**, pp. 1475-1481, 1992.

12. J. Kilday, F. Palmieri, and M. D. Fox, Classifying Mammographic Lesions Using Computerized Image Analysis, *IEEE Transaction on Medical Imaging*, Vol. **12**, No. 4, pp. 664-669, Dec. 1993.

13. Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms, *Acad. Radiol.*, **5**, pp. 155-168, 1998.

14. B. Sahiner, H. P. Chan, N. Petrick, L. M. Hadjiiski, M. A. Helvie and S. Paquerault, "Active contour models for segmentation and characterization of mammographic masses," *Madison, WI: Medical Physics Publishing, The 5th International Workshop on Digital Mammography*, Toronto, Canada, Proc. IWDM-2000, (in press), 2000.

15. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics* (submitted), 2000.

16. S.S. Rao, *Optimization: Theory and Applications*", Wiley Eastern Limited, 1979.

17. F.A. Lootsma, (ed). *Numerical methods for non-linear optimization*, Academic Press, New York, 1972.

18. M. M. Galloway, Texture Analysis Using Gray Level Run Length, *Comput. Graph. Image Process.* **4**, pp. 172-179, 1975.

19. M. J. Norusis, SPSS Professional Statistics 6.1 (SPSS Inc., Chicago, 1993).

20. M. M. Tatsuoka, Multivariate Analysis, Techniques for Educational and Psychological Research (Macmillan, New York, 1988).

21. C. E. Metz, ROC methodology in radiographic imaging, *Invest. Radiol.*, **21**, pp. 720-733, 1986.

22. C. E. Metz, J. H. Shen, and B. A. Herman, New Methods for Estimating a Binomial ROC Curve From Continuously Distributed Test Results, *presented at the 1990 Annual Meeting of the American Statistical Association, Anahaim, CA*, 1990.